Proceedings of SALT 30: 206–226, 2020

Determiners are "conservative" because their meanings are not relations: evidence from verification*

Tyler Knowlton
University of Maryland

Paul Pietroski
Rutgers University

Alexander Williams
University of Maryland

Justin Halberda
Johns Hopkins University

Jeffrey Lidz
University of Maryland

Abstract Quantificational determiners have meanings that are "conservative" in the following sense: in sentences, repeating a determiner's internal argument within its external argument is logically insignificant. Using a verification task to probe which sets (or properties) of entities are represented when participants evaluate sentences, we test the predictions of three potential explanations for the cross-linguistic yet substantive conservativity constraint. According to "lexical restriction" views, words like *every* express relations that are exhibited by pairs of sets, but only some of these relations can be expressed with determiners. An "interface filtering" view retains the relational conception of determiner meanings, while replacing appeal to lexical filters (on relations of the relevant type) with special rules for interpreting the combination of a quantificational expression (Det NP) with its syntactic context and a ban on meanings that lead to triviality. The contrasting idea of "ordered predication" is that determiners don't express genuine relations. Instead, the second argument provides the scope of a monadic quantifier, while the first argument selects the domain for that quantifier. On this view, a determiner's two arguments each have a different logical status, suggesting that they might have a different psychological status as well. We find evidence that this is the case: When evaluating sentences like *every big circle is blue*, participants mentally group the things specified by the determiner's first argument (e.g., the big circles) but not the things specified by the second argument (e.g., the blue things) or the intersection of both (e.g., the big blue circles). These results suggest that the phenomenon of conservativity is due to ordered predication.

Keywords: conservativity, quantification, verification, psychosemantics

* We thank Simon Chervenak for help collecting the data, Zoe Ovans for valuable feedback on previous drafts, and Valentine Hacquard, Norbert Hornstein, Nico Arlotti, and audiences at University of Pennsylvania, University of Maryland, CUNY2020, ELM1, and SALT30 for helpful discussion. Funding was provided by the National Science Foundation (#BCS-2017525 and NRT #1449815).

1 Explaining conservativity

1.1 Describing the phenomenon

Given some circles, every one of them is blue if and only if every one of them is a blue circle. Correspondingly, the meanings of the English sentences (1a) and (1b) are logically equivalent.

(1) a. Every circle is blue.
   b. Every circle is a blue circle.

Interestingly, replacing *every* with any other determiner (e.g., *some*, *most*, *the*, *no*) will preserve the same logical equivalence between (1a) and (1b).

But it's easy to imagine determiners that would violate this pattern, call them "non-conservative determiners." For example, suppose *equi* meant "equal in number" so that (2a) means that there's the same number of circles and blue things and (2b) means that there's the same number of circles and blue circles. These sentences fail to be logically equivalent (imagine one blue circle, one red circle, and one blue square; (2a) can be used to describe this situation while (2b) cannot).

(2) a. Equi circles are blue.
   ≈ *the circles are equinumerous with the blue things*
   b. Equi circles are blue circles.
   ≈ *the circles are equinumerous with the blue circles*

To take another example, suppose *yreve* meant "includes" so that (3a) and (3b) have the meanings given below. To a first approximation, this is the meaning *only* would have if it were a determiner.[1] (3a) and (3b) also fail to be equivalent (for example, (3a) cannot be used to describe a situation where there is one blue circle and one blue square, but (3b) can).

(3) a. Yreve circle is blue.
   ≈ *the circles include all blue things*
   b. Yreve circle is a blue circle.
   ≈ *the circles include all blue circles*

1 We assume that *only* is a focus operator, even in *only (the) circles are blue*, given its relatively free distribution; see Herburger 2001 for discussion and comparison with *even*. Note that unlike a determiner, *only* can be added at any point in a sentence like *the cat thought that the dog found it*. Moreover, *only* is focus-sensitive in that focus matters for the truth-conditions. For example, compare *Students only ordered coffee*$_F$ (they didn't also order tea, soda, etc.) and *Students only ordered*$_F$ *coffee* (they didn't also make it, purchase it, etc.).

These hypothetical words (and many others) are simple and would potentially be communicatively useful, so it's striking that English and other languages lack such determiners. Instead, languages only have determiners that obey the "conservativity constraint": an instance of (4a), where PRED can be a verbal or adjectival predicate, is logically equivalent to the corresponding instance of (4b), allowing variations in morphology and word order. If (4a) is true, (4b) will be true, and vice versa.

(4) a. [[DET NP] PRED]
b. [[DET NP] [be NP that PRED]]

Accordingly, when verifying sentences with conservative determiners, like those in (5), one need not look beyond the circles and how they are colored. Any other things – e.g., salient squares or triangles – are irrelevant.[2]

(5) {Some / Most / Every / The / DET} circles are blue.

Only things that satisfy the first / NP / internal argument of a determiner matter for the truth of the sentence.[3] And given the independent evidence that children can't acquire non-conservative determiners despite being able to acquire novel conservative determiners in the same experimental context (Hunter & Lidz 2013; though see also Spenader & de Villiers 2019), the conservativity constraint seems to be a symptom of how the language faculty operates. Non-conservative determiners are conceivable but not possible lexical items for humans. Our semantic theory should aim to explain this constraint.

1.2 Restricted versus relational quantification

The logical role of the first argument invites the following analogy: the noun *circles* seems to function as the restrictor in a restricted quantifier. One way to capture this formally is to take (6a) to be a sentence of a Tarskian language that is satisfied by certain sequences of assignments of values to variables. We can say that a sequence σ satisfies (6a) if and only if σ assigns x to something blue. Then (6b) is a new sentence formed by adding the prefix $\forall x$. It is satisfied by σ if and only if each x-variant of σ – i.e., each sequence that is like σ except perhaps with regard to what σ assigns to the variable x – assigns x to something blue.

2 *Many* is a potential counterexample to this generalization given "reverse proportional" readings of sentences like *many Scandinavians have won the Nobel prize in literature* (Westerståhl 1985). But Romero (2015) argues that *many* is a gradable adjective that decomposes into a conservative determiner and a degree operator.

3 This phenomenon has been described in various ways. Barwise & Cooper (1981) say that determiners "live on" their internal arguments. Higginbotham & May (1981) describe determiners as "intersective." Keenan & Stavi (1986) say that "determiners are always interpreted by *conservative* functions."

(6) a. $BLUE(x)$
b. $\forall x[BLUE(x)]$
c. $\forall x : CIRCLE(x)[BLUE(x)]$

We can then imagine a restricted variant (6c), that is satisfied by σ if and only if each x-variant of σ that assigns a circle to x also satisfies the open sentence (6a). This confines the sequences to be considered to just those in which x is assigned a circle, in effect restricting the domain against which (6b) is evaluated to just the circles.

In general, if Q is a restricted quantifier, then (7a) is equivalent to (7b), since repeating the restrictor as a conjunct in the scope of Q is logically inert. In this respect, phrases formed by combining determiners with internal arguments (e.g., *every circle*) are like restricted quantifiers.

(7) a. $Qx : Rx(Sx)$
b. $Qx : Rx(Rx \,\&\, Sx)$

Determiners have also been represented as expressing relations between sets, or their characteristic functions (Barwise & Cooper 1981; Keenan & Stavi 1986). For example, *most* in *most circles are blue* can be said to express the dyadic relation in (8), which obtains if and only if the number of things that are elements of both sets is greater than the number of things that are elements of the set of circles but not the set of blue things.

(8) $MOST_x[BLUE(x), CIRCLE(x)]$

Thinking in relational terms invites a useful description of the conservativity constraint: determiners only express relations for which (9) holds.

(9) Necessarily, for any sets X and Y, $[R(X,Y) \equiv R(X \cap Y, Y)]$[4]

But as Westerståhl (2019) shows, the overtly relational (8) is logically equivalent to the non-relational (10) (read as "relativized to the set of circles, most things are blue"). In (10), a monadic quantifier is evaluated in a universe restricted to circles.

(10) $MOST_x[BLUE(x)] \upharpoonright CIRCLES$

4 This is sometimes written as $R(A,B) \equiv R(A, A \cap B)$. Here we follow the convention that an ordered pair $< X,Y >$ can be identified with the set $\{X, \{X,Y\}\}$, in which Y is the internal element. For example, *Brutus saw Caesar*, where *Caesar* is the internal argument of the verb, is usually regimented as $SAW(BRUTUS, CAESAR)$. Likewise, in (9), X is the external argument of the relation and Y is the internal argument. This means that *every circle is blue*, where *circle* is the internal argument of the determiner, corresponds to $SUPERSET(BLUE, CIRCLE)$ instead of the more traditional $SUBSET(CIRCLE, BLUE)$.

This restates our earlier observation about restricted quantification. If determiners express quantifiers over a universe restricted by their internal argument, conservativity is a logical consequence

These distinct analyses – in terms of restricted monadic quantifiers versus genuinely relational quantifiers – thus suggest different potential explanations for the conservativity constraint. In this paper, we consider two proposals that start with the assumption that determiners express relations and explain conservativity by filtering out the problematic ones, either through stipulation (section 1.3) or with the help of a certain interpretation of quantifier raising and a filter on trivial meanings (section 1.4). We contrast these proposals against an "ordered predication" view that rejects relational determiner meanings in favor of the idea that the two arguments have a different logical status: the first argument restricts the domain of evaluation and the second argument supplies the scope of the monadic quantifier (section 1.5). On this view, conservativity is a consequence of the logical role played by the first argument.

Given an independently plausible linking hypothesis about how meaning and verification are related (section 2), these views make different predictions about how participants will treat the first and second arguments of a determiner. These predictions are the focus of our experiments (section 3). Our results favor ordered predication. To preview: when participants evaluate a statement like *every big circle is blue*, they encode the cardinality of the set denoted by the first argument (*big circles*) but not the cardinalities of the sets denoted by the second argument (*blue things*) or the intersection of both arguments (*big blue circles*). This suggests that participants only mentally group the extension of the first argument, which would be surprising given a relational conception of determiner meanings.

1.3 Lexical restriction

Assuming that quantificational determiners express relations, one straightforward approach to explaining conservativity would be to forbid non-conservative relations. Suppose that *every* expresses the superset relation, for example (see note 4). Then one could stipulate that $SUPERSET$ is a relation that can be lexicalized but $SUBSET$ – the meaning of the hypothetical non-conservative *yreve* in (3a) – is not.

Keenan & Stavi (1986) offer one version of this approach. They combine a store of "basic" relations that are conservative (e.g., relational analogues of the Aristotelian quantifiers) with some conservativity-preserving operations for characterizing other relations like those corresponding to *most* and *seventeen*.

Whatever the specifics, though, this lexical restriction view raises questions. Given a fundamentally relational conception of quantification, why should $SUPERSET$ be one of the basic relations, while the non-conservative relations in the immediate conceptual vicinity – including $PROPER\text{-}INCLUSION$ and $SUBSET$ – are not?

And if allowing for improper inclusion (a.k.a. identity) can be motivated, why isn't $IDENTITY$ one of the basic relations? Similarly, even if the operations needed to introduce *most* and/or *seventeen* are motivated, why is $EQUINUMEROSITY$ (which the hypothetical *equi* in (2a) might be said to express) not a candidate for being expressed by an atomic determiner? After all, the notion of one-to-one correspondence, which is available to infants (Wynn 1992; Feigenson 2005), seems fundamental to the "numeric" quantifiers.

So prima facie, this first way of accounting for the phenomenon of conservativity amounts to a redescription of the explanandum in terms of the idea that determiners are special cases of Generalized Quantifiers, namely, the ones that humans can lexicalize. Still, even if this seems unsatisfying, it might be true.

1.4 Interface filtering

Romoli (2015) offers an alternative that provides a potentially more motivated way of explaining the conservativity constraint while maintaining the idea that determiners express set-theoretic relations. Building on suggestions from Chierchia (1995), Fox (2002), and Sportiche (2005), Romoli appeals to the syntax that underpins quantifier raising, along with an ancillary hypothesis about traces of displacement, as a way of ruling out determiners like *equi* and *yreve* (see (2a) and (3a)).

The basic idea is that a sentence like (11a) has a logical form like (11b) after quantifier raising. The second instance of *every* is then converted to an instance of *the*, as in (11c), meaning that while the copy of *circles* gets interpreted, the copy of *every* does not. If this second instance of the first argument *circles* in the verb phrase gets conjoined with the predicate, (11a) has the meaning in (11d).

(11) a. Every circle is blue.
b. [every circle [every circle is blue]]
c. [every circle [~~the circle~~ is blue]]
d. $BLUE\text{-}THINGS \cap CIRCLES \supseteq CIRCLES$

In (11d), *every* still expresses the same relation, but it's relating the circles and the set of blue circles instead of the circles and the set of blue things. This change doesn't affect the truth-conditions (if it did, $SUPERSET$ would fail to be a conservative relation as per the definition in (9)).

But how are non-conservative determiners ruled out? First, consider *equi*. Given quantifier raising and trace conversion (from *equi circle* to *the circle*), a sentence with *equi* would have the logical form in (12c) and the interpretation in (12d).

(12) a. Equi circle is blue.

b. [equi circle [equi circle is blue]]

c. [equi circle [~~the circle~~ is blue]]

d. $BLUE\text{-}THINGS \cap CIRCLES = CIRCLES$

But (12d) is truth-conditionally equivalent to (11d). Because the syntax causes there to always be a copy of the first argument in the verb phrase, then *every* might in fact express $IDENTITY$, not $SUPERSET$. But even so, any sentence with *every* would result in a conservative meaning. The problematic (2a) would never arise. In other words: *equi* might exist, but if it did, any time it got used in a sentence, it would be truth-conditionally equivalent to the same sentence with *every*. And as Romoli (2015) shows, this line of thinking holds for a subset of the would-be non-conservative determiners: given their syntax, sentences with them are truth-conditionally equivalent to sentences with a conservative determiner.

For another class of potential non-conservative determiners, like *yreve*, an extra ingredient is needed. The interpretation after quantifier raising and trace conversion – (13d) – is not truth-conditionally equivalent to a sentence with any extant conservative determiner. In fact, (13d) will always be true, as the circles will always be a superset of or be identical to the blue circles.

(13) a. Yreve circle is blue.

b. [yreve circle [yreve circle is blue]]

c. [yreve circle [~~the circle~~ is blue]]

d. $BLUE\text{-}THINGS \cap CIRCLES \subseteq CIRCLES$

So, a filter on trivial meanings (e.g., Gajewski 2002; Fox & Hackl 2006) – sentences that are tautologies or contradictions for any substitution of lexical content – is posited to rule out *yreve*. If it existed, any sentence with *yreve* would be declared ungrammatical given such a filter.

1.5 Ordered predication

The non-relational ordered predication view also assumes raising for quantificational determiner phrases, even those in subject position, so that the relevant syntax for (14a) is (14b).

(14) a. Every circle is blue.

b. $[_{S'}$ [every circle]$_x$ $[_S$ $\cancel{\text{[every circle]}_x}$ is blue]]

In contrast with interface filtering though, here the indexed trace (or lower copy) of the raised DP is treated as a variable in the usual way. Then, we can think of the embedded clause as an open sentence akin to it_x *is blue*. This open sentence is satisfied by an assignment of values to variables if and only if the thing assigned to the variable is blue. The whole sentence in (14a) is true, relative to an arbitrary assignment A, just in case every circle, x, is such that the embedded open sentence is true relative to the variant of A that assigns x to the indexed variable; see Pietroski 2018.

Alternatively, instead of thinking of the first argument as restricting the assignments with respect to which the open sentence is evaluated, we could think of it as restricting the domain of evaluation. Using the "hook" notation given above (section 1.2), we can read (15) as "relativized to the set of circles, everything is blue."

(15) $\forall x[BLUE(x)] \upharpoonright CIRCLES$

Either way, the crucial point for our purposes is that the domain with respect to which the quantified expression is evaluated is restricted by the first (nominal) argument. The determiner describes how the further condition supplied by the second argument (*is blue*) applies to the members of the restricted domain (the circles). In the case of *every*, for example, it applies exhaustively.

All conservative determiners can be stated in this way, even proportional quantifiers like *most* (though see Pietroski, Lidz, Hunter & Halberda 2009 and Lidz, Pietroski, Halberda & Hunter 2011 on what it means for a predicate to apply to *most* things in a restricted domain). But crucially, non-conservative determiners are not stateable (Westerståhl 2019).

For example, consider *equi* in (2a). The intended meaning is that the circles are equinumerous with the blue things. But this cannot be stated in terms of how the predicate *is blue* applies to the circles. One way or another, the cardinality of the blue things needs to be taken into account, and the blue things are outside of the restricted domain. The same is true for *yreve* in (3a) and any other potential determiner that requires making reference to the extension of the second argument. So conservativity is a logical consequence of ordered predication.

2 Testing predictions of the three views

2.1 Linking hypothesis

Importantly, each of the three proposals makes a different prediction about the semantic representation for sentences with conservative determiners. In this paper, we use *every* as a case-study. Consider the meaning of (16a) according to each view:

(16) a. Every N is P.

b. $P \supseteq N$ (lexical restriction)

c. $P \cap N \supseteq N$ (interface filtering)

d. $\forall x[P(x)] \upharpoonright N$ (ordered predication)

We adopt the linking hypothesis from Lidz et al. (2011): The verification procedures employed in understanding a declarative sentence are biased towards algorithms that directly compute the relations and operations expressed by the semantic representation of that sentence.

In a semantic representation like (16b), two independent sets are explicitly related (those denoted by the first and second arguments). Likewise, in (16c) two sets are related (one denoted by the first argument and the other by the intersection of the first and second arguments). If either of these is the meaning of *every N is P*, then, all else equal, the most natural thing for participants to do when trying to verify that meaning would be to represent and relate the two specified sets. The visual system can represent up to three sets in parallel with no loss of acuity (Halberda, Sires & Feigenson 2006). Intuitively, adults (and children) routinely do represent multiple sets when evaluating sentences like *there are more blue dots than yellow dots*. So there is no independent constraint against representing and relating both sets.

But in (16d), the first and second arguments are not treated on a par logically. It is less clear which items participants should mentally group, if any (this may depend on whether the restriction is better thought of in terms of restricting assignments or restricting the universe of evaluation). But given the logical asymmetry between the two arguments, (16d) predicts a corresponding psychological asymmetry in the way that participants treat their extensions.

Our experiments thus probe which set(s) participants represent when evaluating quantificational sentences. As a proxy for whether participants represented a set, we probe their memory for the set property cardinality (see Knowlton, Pietroski, Halberda & Lidz under review). Intuitively, cardinality is a property of groups much like average size or center of mass. When people represent a group of individuals as a group – sometimes called an ensemble representation – they abstract away from individual properties and instead encode these sorts of set summary statistics (e.g., Ariely 2001; Halberda et al. 2006; Burr & Ross 2008; Alvarez 2011). So, if participants represented a set as such, they should have a better estimate of its cardinality than if they did not.

2.2 Experimental design

Participants first saw a statement like *every big circle is blue*. Sizes (big, medium, small) and colors (blue, red, yellow) were randomized. After pressing "spacebar", participants were briefly shown an image of between 24 and 48 circles on a grey background (Fig. 1). Medium circles had grey holes in the middle, to make them

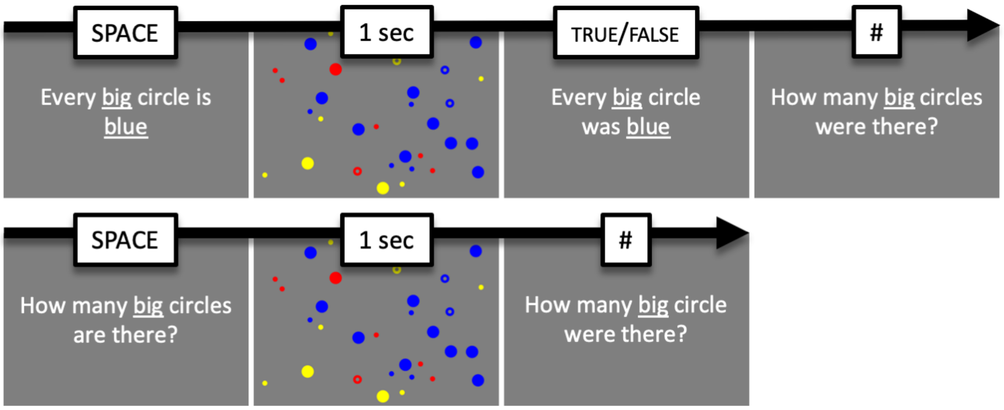

Figure 1 Participants completed a sentence verification task (top) and a baseline number-knowledge task using the same image types (bottom).

more distinguishable from the other two sizes (Chen 1982; Chen 2005). Participants' task was to judge the statements as true or false, relative to the picture. They responded by pressing "J" or "F" on their keyboard.

After giving their true/false response, participants were given a follow-up cardinality question (e.g., *how many big circles were there?*) and were asked to type their guess before moving onto the next trial. Some questions probed the set denoted by size (big, in this example; but sometimes small or medium), others probed the set denoted by color (blue, in this example; but sometimes red or yellow), and others probed the intersection of size and color (big blue, in this example; but any combination of size and color was possible). Each participant saw 18 trials in total.

Each participant also first completed a 15-trial baseline number task (Fig. 1) based on Halberda et al. (2006). In this task, there were no sentences to evaluate, and the cardinality question was shown prior to the dot display. Performance on this task represents the best possible cardinality estimates the visual system will allow given these displays and a one second viewing limit. It also controls for any differences in difficulty (e.g., perhaps in these particular displays it is easier to visually group circles based on their color than on their size).

Each of the six experiments in the following section used this basic design. Experiment 1 tested sentences with a first argument defined by size and a second argument defined by color, as in *every big circle is blue*. We find that while participants know the cardinality of the set denoted by *every*'s first argument (*big circles*) as well as they know the cardinality of sets defined by size at baseline, they are significantly worse than baseline performance when asked to recall the cardinality of the set denoted by *every*'s second argument (*blue circles*) or the cardinality of the set denoted by the intersection of both arguments (*big blue circles*). This suggests that participants only mentally group the extension of *every*'s first argument. The five subsequent experiments control for various alternative explanations.

3 Results

3.1 Experiment 1: *every big circle is blue*

71 participants were recruited on Amazon Mechanical Turk. Participants were excluded from further analysis if they failed an English proficiency screener (18), performed at chance or below on the true/false portion (2), or took longer than 5 seconds, on average, to respond to cardinality questions (3). This left 48 participants.[5]

Responses to cardinality questions in both the baseline and sentence-verification tasks were fit with the standard psychophysical model of number estimation (Stevens 1964; Krueger 1984; Odic, Im, Eisinger, Ly & Halberda 2016). This model allows participants' accuracy on number questions (i.e., the average proximity to the true cardinality) to be described by a single parameter (β in the equation $y = \alpha x^{\beta}$, where y is the numerical response, x is the true cardinality, and α is a scaling factor). This measure of accuracy on the cardinality questions following sentence-verification was then subtracted from the same measure of accuracy obtained from the baseline trials, creating a single accuracy difference score for each set tested. Error around this difference score was computed by taking the square root of the sum of squared errors around both accuracies.

Using these two values, we conducted Wald tests to compare the difference score for each subset probed (size, color, and intersection) against the null hypothesis of a difference score of 0. A significant result thus indicates that participants were worse than their visual system would allow when asked to give cardinality estimates of that subset after sentence evaluation. If the set is represented following evaluation, then performance should not be statistically worse than baseline. But if the set in question is not represented, performance should be significantly worse.

As mentioned in section 2, each of the three explanations of conservativity makes a different prediction about which set(s) participants should represent during evaluation. Lexical restriction predicts mental grouping of both the first and second arguments (though perhaps not their intersection). Interface filtering predicts mental grouping of both the first argument and the intersection of the first and second arguments (though perhaps not the second argument by itself). Ordered predication predicts that participants will treat the two arguments asymmetrically.

We find that participants know the cardinality of the set denoted by the first argument (*big circles*): their difference score is not significantly lower than 0 ($\chi^2 = 0.007, p = .93$; Fig. 2). In other words, after evaluating a sentence like *every big*

5 For all subsequent experiments, we aimed for a final sample of 48, given these exclusion criteria.

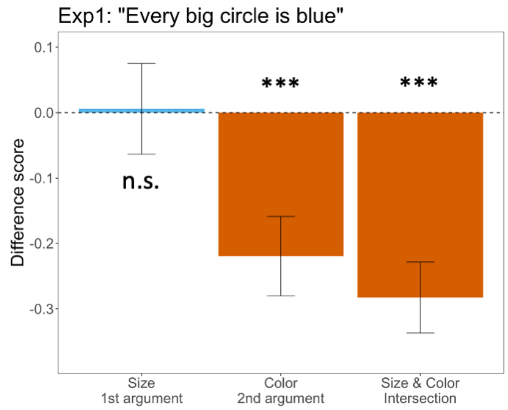

Figure 2 Participants' baseline accuracy minus their accuracy on number-questions after verifying statements like *every big circle is blue*. Here and for subsequent experiments, difference scores when asked about the set denoted by the first argument of the determiner are shaded in blue; difference scores when asked about other sets are shaded in orange.

circle is blue, they seem to know the cardinality of the big circles as well as their visual system will allow; just as well as when they were instructed ahead of time to pay attention to the big circles and estimate their cardinality.

However, participants were significantly worse than their baseline performance when asked about the set denoted by the second argument (*blue circles*) ($\chi^2 = 13.13, p < .001$; Fig. 2) and when asked about the set denoted by the intersection of both arguments (*big blue circles*) ($\chi^2 = 27.09, p < .001$; Fig. 2). This latter result is especially surprising as on half of the trials the statement was true (e.g., every big circle was blue), and in those cases the cardinality of the first argument and the cardinality of the intersection was identical.

This result shows that participants only mentally group the extension of *every*'s first argument. We take this to be very surprising on either of the two relational views (lexical restriction and interface filtering), given that (i) these views hold that the semantic representation of *every big circle is blue* is a relation between two sets and (ii) there is no apparent non-linguistic pressure that would prevent participants from adopting a verification strategy that relies on representing and relating two sets. On the other hand, this result fits nicely with the non-relational ordered predication view: the semantic representation treats both arguments differently, so participants treat their extensions differently as well.

That said, there are a number of alternative factors that might have encouraged participants to represent the first arguments as a set. Some of these will be addressed in the experiments that follow while others will be left for future work.

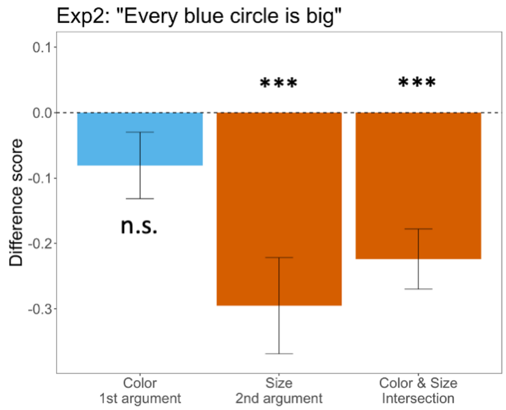

Figure 3 Participants' baseline accuracy minus their accuracy on number-questions after verifying statements like *every blue circle is big*.

First, perhaps there is something special about the fact that *every*'s first argument was based on size instead of color (experiment 2). Second, perhaps participants always represent the extension of NPs that they encounter as sets, regardless of the sentence meaning (experiments 3-4). Third, perhaps participants would have represented the second argument as a set but could not because they had reached their working memory capacity limit by the time they made it to the verb phrase (experiment 5). Lastly, we show that symptoms of the asymmetry can be replicated without even needing to examine participants' responses to the various cardinality questions (experiment 6).

3.2 Experiment 2: *every blue circle is big*

Regarding the first possibility: might there be something special about the quantifier's first argument being specified by a size? Perhaps, for example, participants always represent sets defined by size given displays of this type. If so, then we should expect the set denoted by the mentioned size to be represented even when it does not appear as *every*'s first argument. For that reason, experiment 2 was identical to experiment 1 except that the arguments were inverted, yielding sentences like *every blue circle is big*. On the other hand, if participants represent *every*'s first argument as a set regardless, then even given these "swapped argument" sentences, participants should represent the set described by color but not the set described by size (or by the intersection of both).

This latter prediction was borne out. Participants' estimates of the first argument's cardinality (*blue circles*) following sentence-verification were not significantly different from their estimates during baseline ($\chi^2 = 2.53, p = .11$; Fig. 3).

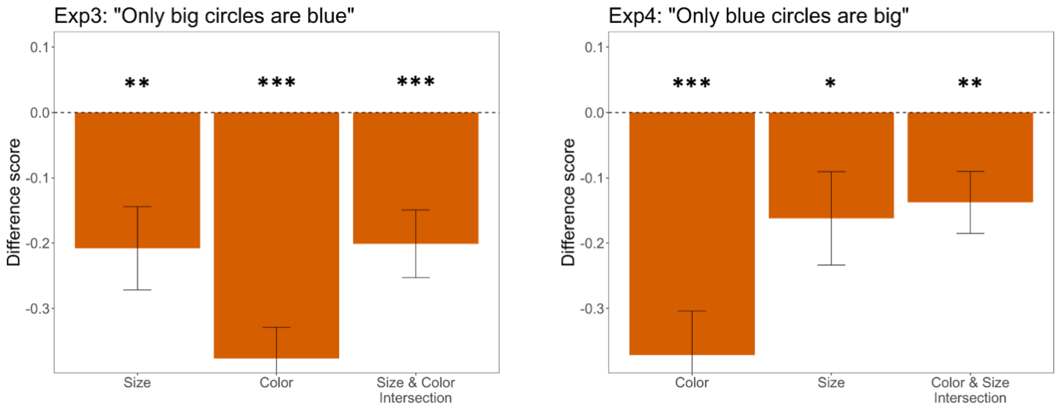

Figure 4 [Left] Participants' baseline accuracy minus their accuracy on number-questions after verifying statements like *only big circles are big* and [Right] after verifying statements like *only blue circles are big*.

But they did perform significantly worse than baseline when asked about the set denoted by the second argument (*big circles*) ($\chi^2 = 16.16, p < .001$; Fig. 3) or the intersection (*big blue circles*) ($\chi^2 = 23.69, p < .001$; Fig. 3). This confirms that there is something important about being the first argument of the quantifier with respect to triggering set representation.

3.3 Experiments 3 and 4: *only big circles are blue / only blue circles are big*

Another possibility is that participants performed well when asked about the cardinality of the first argument in experiments 1 and 2 because being an NP always triggers set representation. It is worth noting, however, that given one of the relational views, this would mean that something other than the meaning is responsible for triggering set representation (since, on these views, both the NP and VP are treated as sets).

To rule out this possibility empirically, experiments 3 and 4 replace *every* with *only*. Other than that, these experiments are identical to experiments 1 and 2. Because *only* is not a determiner (see note 1), it does not take arguments in the same way as *every*. If being the first argument of a determiner is important for triggering this sort of set representation during verification, then we predict no sets to be represented given these *only*-variants.

Consistent with this prediction, we find that changing *every* to *only* causes participants to perform significantly worse than baseline when asked about any subset. Indeed, they performed significantly worse than baseline in all three conditions after evaluating sentences like *only big circle are blue* (size: $\chi^2 = 10.67, p < .01$; color:

$\chi^2 = 62.08, p < .001$; intersection: $\chi^2 = 14.99, p < .001$; Fig. 4, left) and after evaluating sentences like *only blue circles are big* (color: $\chi^2 = 30.15, p < .001$; size: $\chi^2 = 5.14, p < .05$; intersection: $\chi^2 = 8.34, p < .01$; Fig. 4, right).

This suggests that *every* was particularly relevant in triggering set representation of its first argument in experiments 1 and 2. Merely being an NP is not enough to encourage participants to mentally group the extension.

3.4 Experiment 5: *the blue team painted every big circle*

A remaining alternative explanation is that participants would have mentally represented both arguments (or the first argument and the intersection) as sets, but ran up against their working memory capacity. As noted above, humans can represent three sets in parallel before incurring costs Halberda et al. (2006). One of these is always the superset of all dots, leaving two remaining "slots". When attempting to evaluate *every big circle is blue*, participants might first represent the big circles and then may fill the second "slot" with another set (e.g., perhaps the complement of this set). If so, they would have no "room" to represent the extension of *is blue*, which, by hypothesis, is the set of blue circles.

To control for this possibility, experiment 5 put the second argument into subject position. Participants were told that they had to judge a contest between the blue, red, and yellow teams, each of which were painting circles their color. Then they were asked to judge statements like *the blue team painted every big circle* as true or false. Plausibly, given the experimental context, the extension of *the blue team* is the blue circles. So if being earlier in the sentence is to blame fo the fact that participants mentally group the extension of the first argument in experiments 1 and 2, then participants should mentally group the blue circles in this task.

This prediction is not borne out. Participants' cardinality estimates were significantly worse than baseline when asked about a set denoted by color ($\chi^2 = 46.48, p < .001$; Fig. 5) despite color being named first in the sentence.

Participants also performed significantly worse than baseline when asked about a set denoted by a size ($\chi^2 = 6.11, p < .05$) and when asked about a set denoted by the intersection of both ($\chi^2 = 15.87, p < .001$). Given the results of experiments 1 and 2, it is somewhat surprising that participants in experiment 5 are worse than baseline for size questions. This suggests that being in subject position (or being a sentence topic) might play a role in triggering set representation. Alternatively, this might reflect the overall difficulty of asking the question in this way as opposed to the more straightforward *every big circle is blue*. Future work will aim to distinguish these possibilities experimentally.

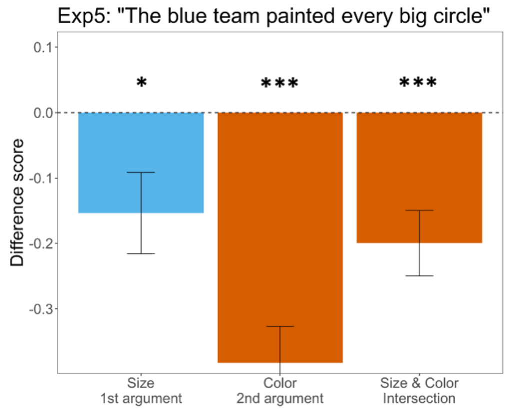

Figure 5 Participants' baseline accuracy minus their accuracy on number-questions after verifying statements like *the blue team painted ever big circle*.

3.5 Experiment 6: *every big circle is blue* (with an "I don't know!" button)

Experiment 6 is identical to experiment 1 except for the addition of a large red "I don't know!" button underneath every cardinality question. The button gave participants a penalty-free option to opt out of any trial they thought they could not accurately answer. They were, however, instructed to use it only as a last resort.

This manipulation provides a replication of the initial experiment using a simple behavioral measure that does not require fitting and interpreting psychophysical models. Given the results of experiments 1 and 2, we predict higher than baseline rates of opting out for questions probing the second argument and the intersection of both arguments. In addition, it offers some insight into whether participants are conscious of their epistemic limitations: do they realize that they often don't have a good estimate of the second argument's or the intersection's cardinality?

The overall rate of "I don't know!" button presses during each portion of the task is plotted in Fig. 6 (left). We find that participants are no more likely than baseline to opt out of cardinality questions probing the first argument (*big circles*) ($t_{47} = 0.47, p = .64$). But they are significantly more likely to opt out of questions probing the scope argument (*blue circles*) ($t_{47} = 2.91, p < .01$) and for those probing the intersection (*big blue circles*) ($t_{47} = 3.07, p < .01$).

As in experiments 1 and 2, this suggests that participants only represent the first argument of the determiner as a set. When asked about the first argument's cardinality, they are as confident about their responses as their visual system allows. But when asked similar questions about the second argument's or the intersection's cardinality, they are less confident than expected.

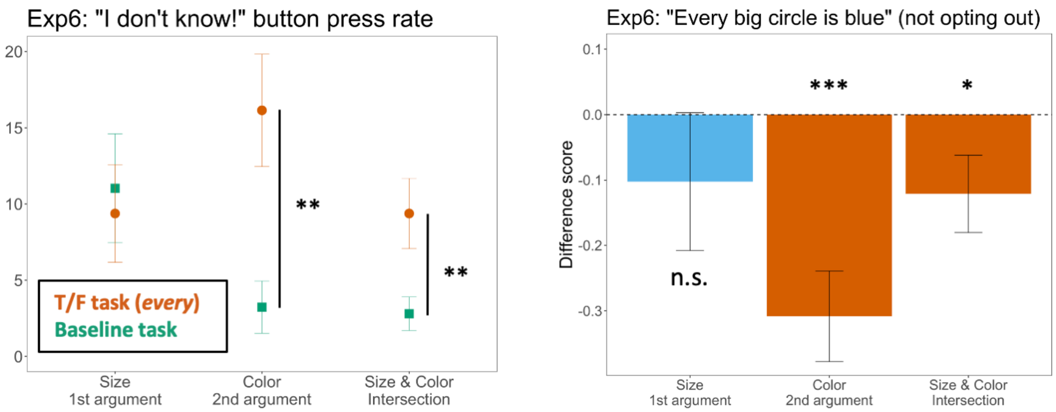

Figure 6 [Left] Participants' average rate of pressing the "I don't know!" button instead of answering the cardinality question during the baseline task (orange) and following sentence-verification (green). [Right] Participants' baseline accuracy minus their accuracy on number-questions for the trials that they did not opt out of.

Moreover, when considering only the trials on which participants did not opt out (i.e., trials on which they gave a numeric answer instead of pressing the "I don't know!" button), we find the same effect as in experiments 1 and 2 (Fig. 6, right). Namely, participants were significantly worse than baseline on questions probing the second argument ($\chi^2 = 19.83, p < .001$) and the intersection of both arguments ($\chi^2 = 4.19, p < .05$), but their performance was not significantly different from baseline on questions probing the cardinality of the first argument ($\chi^2 = 0.94, p = .33$). This suggests that while participants are somewhat aware of their epistemic limitations, they do not have conscious access to its full extent.

4 Conclusions and future directions

4.1 What triggers set representation?

In our view, these results support a non-relational view of determiner meanings. The reason that a sentence like *every big circle is blue* encourages representation of the big circles as a set – and not the blue ones or the big blue ones – is because its meaning is more in line with (16d) than (16b) or (16c).

To be sure, meaning is not the only factor that carries weight in determining how participants will approach a given scene in a sentence-verification task. Details of the visual display, for example, can bias participants to represent one group of circles as a set but not another. But because the current results are compared against a baseline and because the predictions hold when the arguments are flipped, visual properties of the display are not likely to be at issue here.

Still, there may be other details of the sentences used that are responsible for triggering set representation. As noted, sentence topicality may be of particular interest, as being the topic may play a role in encouraging participants to mentally group individuals.

Another possibility is that participants are, like Barwise & Cooper (1981) suggested, only choosing to represent the first argument of *every* as a set during verification because sentences with *every* have conservative meanings and can thus be verified without looking beyond the extension of the first argument, as noted in section 1. It's a logical possibility that participants understand *every* to express a relation between two sets, but nonetheless only mentally group the extension of the first argument when evaluating sentences with *every*. As mentioned though, given that participants' visual systems represent three sets in parallel at no cost to performance (Halberda et al. 2006), there is no obvious sense in which representing only one is an easier or otherwise superior strategy.

4.2 What about other determiners?

Crucially, the non-relational ordered predication explanation of conservativity does not predict that all determiners will lead to identical performance when it comes to triggering set representation. Some determiners may be represented in a completely first-order way that eschews sets or any other notion of grouping the satisfiers of a predicate together. In other work using a similar paradigm, we argue that *each* is one such determiner (Knowlton et al. under review). Other determiners, like *most*, may lead to representation of more than one set (e.g., Lidz et al. 2011).

What is predicted by ordered predication is that a determiner's second argument (like *is blue*) will not trigger group representation. Instead, the second argument is treated as an open sentence that applies to the members of a restricted domain in the way specified by the determiner (e.g., exhaustively in the case of *every*; proportionally in the case of *most*; with a cardinality restriction in the case of *two*). While the domain may be initially be restricted in a way that requires treating the first argument as a group (as we have shown for *every*) or in a way that implicates only individuals (as we suspect for *each*), the application of the second predicate should never trigger group representation.

In future work, we hope to use similar methods to investigate other determiners and further test this prediction. If our findings generalize beyond *every*, ordered predication offers a ready explanation of our participants' behavior and a simple account of the conservativity constraint: no determiner expresses a non-conservative relation because no determiner expresses a relation in the first place.

References

Alvarez, George A. 2011. Representing multiple objects as an ensemble enhances visual cognition. *Trends in Cognitive Sciences* 15(3). 122–131. doi:10.1016/j.tics.2011.01.003.

Ariely, Dan. 2001. Seeing sets: Representation by statistical properties. *Psychological Science* 12(2). 157–162. doi:10.1111/1467-9280.00327.

Barwise, Jon & Robin Cooper. 1981. Generalized quantifiers and natural language. In *Philosophy, Language, and Artificial Intelligence*, 241–301. Springer. doi:10.1007/978-94-009-2727-8_10.

Burr, David & John Ross. 2008. A visual sense of number. *Current Biology* 18(6). 425–428. doi:10.1016/j.cub.2008.02.052.

Chen, Lin. 1982. Topological structure in visual perception. *Science* 218(4573). 699–700. doi:10.1126/science.7134969.

Chen, Lin. 2005. The topological approach to perceptual organization. *Visual Cognition* 12(4). 553–637. doi:10.1080/13506280444000256.

Chierchia, Gennaro. 1995. Lecture notes.

Feigenson, Lisa. 2005. A double-dissociation in infants' representations of object arrays. *Cognition* 95(3). B37–B48. doi:10.1016/j.cognition.2004.07.006.

Fox, Danny. 2002. Antecedent-contained deletion and the copy theory of movement. *Linguistic Inquiry* 33(1). 63–96. doi:10.1162/002438902317382189.

Fox, Danny & Martin Hackl. 2006. The universal density of measurement. *Linguistics and Philosophy* 29(5). 537–586. doi:10.1007/s10988-006-9004-4.

Gajewski, Jon. 2002. L-analyticity and natural language. *Manuscript, MIT* http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.488.3195.

Halberda, Justin, Sean F Sires & Lisa Feigenson. 2006. Multiple spatially overlapping sets can be enumerated in parallel. *Psychological Science* 17(7). 572–576. doi:10.1111/j.1467-9280.2006.01746.x.

Herburger, Elena. 2001. The negative concord puzzle revisited. *Natural Language Semantics* 9(3). 289–333. doi:10.1023/A:1014205526722.

Higginbotham, James & Robert May. 1981. Questions, quantifiers and crossing. *The Linguistic Review* 1(1). 41–80. doi:10.1515/tlir.1981.1.1.41.

Hunter, Tim & Jeffrey Lidz. 2013. Conservativity and learnability of determiners. *Journal of Semantics* 30(3). 315–334. doi:doi.org/10.1093/jos/ffs014.

Keenan, Edward L & Jonathan Stavi. 1986. A semantic characterization of natural language determiners. *Linguistics and Philosophy* 9(3). 253–326. https://www.jstor.org/stable/25001246.

Knowlton, Tyler, Paul Pietroski, Justin Halberda & Jeffrey Lidz. under review. The mental representation of universal quantifiers https://ling.auf.net/lingbuzz/004486.

Krueger, Lester E. 1984. Perceived numerosity: A comparison of magnitude production, magnitude estimation, and discrimination judgments. *Perception & Psychophysics* 35(6). 536–542. doi:10.3758/BF03205949.

Lidz, Jeffrey, Paul Pietroski, Justin Halberda & Tim Hunter. 2011. Interface transparency and the psychosemantics of *Most*. *Natural Language Semantics* 19(3). 227–256. doi:10.1007/s11050-010-9062-6.

Odic, Darko, Hee Yeon Im, Robert Eisinger, Ryan Ly & Justin Halberda. 2016. Psimle: A maximum-likelihood estimation approach to estimating psychophysical scaling and variability more reliably, efficiently, and flexibly. *Behavior Research Methods* 48(2). 445–462. doi:10.3758/s13428-015-0600-5.

Pietroski, Paul, Jeffrey Lidz, Tim Hunter & Justin Halberda. 2009. The meaning of 'most': Semantics, numerosity and psychology. *Mind & Language* 24(5). 554–585. doi:10.1111/j.1468-0017.2009.01374.x.

Pietroski, Paul M. 2018. *Conjoining Meanings: Semantics Without Truth Values*. Oxford University Press.

Romero, Maribel. 2015. The conservativity of *many*. In *20th Amsterdam Colloquium*, 20–29. https://ling.sprachwiss.uni-konstanz.de/pages/home/romero/papers/romero_v8-AC20.pdf.

Romoli, Jacopo. 2015. A structural account of conservativity. *Semantics-Syntax Interface* 2(1). 28–57. https://pure.ulster.ac.uk/en/publications/a-structural-account-of-conservativity-3.

Spenader, Jennifer & Jill de Villiers. 2019. Are conservative quantifiers easier to learn? evidence from novel quantifier experiments. In *22nd Amsterdam Colloquium*, http://hdl.handle.net/11370/79051821-bdba-4168-8214-5cf1033b8451.

Sportiche, Dominique. 2005. Division of labor between merge and move: Strict locality of selection and apparent reconstruction paradoxes. *Manuscript, UCLA* https://ling.auf.net/lingbuzz/000163/current.pdf.

Stevens, SS. 1964. Concerning the psychophysical power law. *Quarterly Journal of Experimental Psychology* 16(4). 383–385. doi:10.1080/17470216408416398.

Westerståhl, Dag. 1985. Logical constants in quantifier languages. *Linguistics and Philosophy* 8(4). 387–413. doi:10.1007/BF00637410.

Westerståhl, Dag. 2019. Generalized quantifiers. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University winter 2019 edn. https://plato.stanford.edu/archives/win2019/entries/generalized-quantifiers.

Wynn, Karen. 1992. Children's acquisition of the number words and the counting system. *Cognitive Psychology* 24(2). 220–251. doi:10.1016/0010-0285(92)90008-P.

Tyler Knowlton
1413H Marie Mount Hall
College Park, MD 20742
tzknowlt@umd.edu

Paul Pietroski
106 Somerset Street
New Brunswick, NJ 08901
paul.pietroski@rutgers.edu

Alexander Williams
1401D Marie Mount Hall
College Park, MD 20742
alxndrw@umd.edu

Justin Halberda
231 Ames Hall
Baltimore, MD 21218
halberda@jhu.edu

Jeffrey Lidz
1417 Marie Mount Hall
College Park, MD 20742
jlidz@umd.edu