Cognition 168 (2017) 294–311

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Original Articles

Resolving uncertainty in plural predication ☆

Gregory Scontras [a,*], Noah D. Goodman [b]

[a] University of California, Irvine, United States
[b] Stanford University, United States

ARTICLE INFO

Article history:
Received 19 October 2015
Revised 30 June 2017
Accepted 5 July 2017

Keywords:
Plurality
Distributive vs. collective interpretations
Ambiguity resolution
Rational Speech Act models

ABSTRACT

Plural predications (e.g., "the boxes are heavy") are common sources of ambiguity in everyday language, allowing both distributive and collective interpretations (e.g., the boxes each are heavy vs. the boxes together are heavy). This paper investigates the role of context in the disambiguation of plural predication. We address the key phenomenon of "stubborn distributivity," whereby certain predicates (e.g., *big*, *tall*) are claimed to lack collective interpretations altogether. We first validate a new methodology for measuring the interpretation of plural predications. Using this method, we then analyze naturally-occurring plural predications from corpora. We find a role of context, but no evidence of a distinct class of predicates that resists collective interpretations. We further explore the role of context in our final experiments, showing that both the predictability of properties and the knowledgeability of the speaker affect disambiguation. This suggests a pragmatic account of how ambiguous plural predications are interpreted. In particular, stubbornly distributive predicates are so because the collective properties they name are unpredictable, or unstable, in most contexts; this unpredictability results in a noisy collective interpretation, something speakers and listeners recognize as ineffective for communicating efficiently about their world. We formalize the pragmatics of utterance disambiguation within the Bayesian Rational Speech Act framework.

1. Introduction

The semantics and pragmatics of predicating properties to singular individuals (e.g., "The box is big") is already a complicated endeavor. What does it mean for a box to be big, or tall, or heavy? How do we represent its size or height or weight? Recent answers to these sorts of questions highlight the central role of context in the calculus of meaning: what properties get ascribed depends crucially on the context of predication (cf. tall for a boy vs. tall for a basketball player; e.g., Kamp & Partee, 1995; Kennedy, 1999; Lassiter & Goodman, 2013). The complexity associated with answering these questions increases once we introduce pluralities into the mix. Particularly puzzling is the distributive vs. collective ambiguity of plural predication (e.g., Link, 1983, 1987, 1998; Scha, 1984; Landman, 1989a, 1989b, 1996; Lasersohn, 1988, 1990, 1995, 1998; Schwarzschild, 1994, 1996):

(1)	The boxes are heavy.	
	a. The boxes each are heavy.	DISTRIBUTIVE
	b. The boxes together are heavy.	COLLECTIVE

When hearing (1), we may infer that each box on its own counts as heavy, as in (1a), or only that their total weight is heavy, as in (1b). Under a collective interpretation, many light boxes may together count as heavy (Scha, 1984). How is the interpretation determined?

To help answer this question, we consider the well-documented yet poorly-understood phenomenon of "stubborn distributivity" (Quine, 1960; Schwarzschild, 2011; Syrett, 2015; Vázquez Rojas Maldonado, 2012; Zhang, 2013), whereby some gradable predicates do not admit collective construals (e.g., *big*, *tall*, *round*), in contrast to those that do (e.g., *heavy*, *expensive*, *voluminous*). For instance, a speaker is hard-pressed to use the sentence in (2) to say of the boxes that their total size is large.

(2)	The boxes are big.	
	a. The boxes each are big.	DISTRIBUTIVE
	b. The boxes together are big.	✗COLLECTIVE

☆ This work was supported in part by a James S. McDonnell Foundation Scholar Award (to N.D.G.); Office of Naval Research Grant N000141310788 (to N.D.G.); and the DARPA Probabilistic Programming for Advanced Machine Learning (PPAML) program, agreement number FA8750-14-2-0009 (to N.D.G.).

* Corresponding author.
E-mail address: g.scontras@uci.edu (G. Scontras).
URL: http://linguistics.uci.edu/scontras/ (G. Scontras).

http://dx.doi.org/10.1016/j.cognition.2017.07.002
0010-0277/

Size and shape adjectives resist collective interpretations. Schwarzschild (2011) terms them "stubbornly distributive," signaling the lack of collective interpretations. Zhang (2013) labels this class of predicates "delimitable," highlighting their usage to describe physical extents. Supporting the robustness of this phenomenon, Syrett (2015) finds evidence of stubborn distributivity in children as young as 3 years of age.

Collective predication arises when a property is ascribed directly to a plurality, rather than distributed among its members (Link, 1983). Semanticists treat stubborn distributivity as a lexical distinction: stubbornly distributive predicates behave as such because they cannot contain pluralities in their basic denotations (Schwarzschild, 2011; Vázquez Rojas Maldonado, 2012; Zhang, 2013).[1] This explanation suffers two flaws: First, it hypothesizes a categorical distinction that may not hold; in Expt. 2, we find a gradient of collectivity with no clear boundary. Second, it leaves unexplained why some predicates would contain pluralities and others not—is there a deeper explanation in terms of the properties named?

Note that stubbornly distributive predicates like *big* or *tall* contrast with truly distributive predicates like *be a man* or *have blue eyes*, whose meaning requires a distributive interpretation (Link, 1983). Having blue eyes or being a man is a property that necessarily holds of individuals, hence their mandatory distributivity; (collective) size is not such a property, at least not transparently so.[2] There is no conceptual barrier to attributing some size directly to a plurality of individuals. Collective nouns bring this point into focus: if a group of men has blue eyes, then each member must; if a pile of boxes counts as tall, we make no direct claim about individual box heights.

(3) a. The group of men has blue eyes.
b. The pile of boxes is tall.

Rather than roaming the lexicon with name tags specifying their attitudes toward collective interpretations (a state awfully difficult for a child to acquire), something about these predicates or the properties they name must influence or determine their behavior in plural predication. The task at hand is to determine what that thing is.

We suggest that the key underlying notion is the stability or *contextual predictability* of a property when applied to a collection.[3] That is, there are aspects of context that affect the computation of collective properties, over and above those that affect the individual properties; when there is more uncertainty about these aspects of context, the collective property is less stable. For instance, collective *tall* names the vertical extent of a set of objects, which can vary dramatically with their arrangement. In contrast, collective *heavy* names the total weight of a set of objects, which does not vary with arrangement. If the precise arrangement is not in common ground, then the collective use of *tall* would be more likely to mislead a listener than that of *heavy*—the listener and the speaker could be encountering different arrangements. In contrast, the distributive meaning of each predicate is fully specified, independent of arrangement. A listener would thus not expect a cooperative speaker to intend the collective meaning of *tall*.

Here is the general hypothesis: computing the meaning of a collective interpretation is susceptible to a varying amount of noise depending on the contextual predictability of the property at issue. To communicate effectively and avoid possible confusion, collective interpretations will be avoided by speakers in the case of such noisy predicates; distributive interpretations, being less variable, are the safe bet. Listeners know this, hence collective interpretations will become more likely as contextual predictability of the collective property increases. In Section 6, we formalize the notion of contextual predictability of the collective property and its role in the pragmatic disambiguation of the meaning of a plural predication. This account explains the phenomenon of stubborn distributivity as a result of the relative unpredictability of the collective property for predicates like *big* and *tall*. It also predicts that context should matter: in a context where the collective property is more stable, the collective reading should be more felicitous.

For physical properties, the stability of physical arrangements is a key element of context. If arrangements are less likely to change, collective properties become more predictable, and so collective interpretations become more useful and more likely. We test this prediction in Expt. 3, measuring whether collective *big* and *tall* indeed feel more natural for boxes that are always stacked regularly than for jumbled piles of boxes. More generally, some types of objects may have more predictable collective properties than others by virtue of how regularly they actualize in the world. If we could find objects that actualize in regular physical arrangements (e.g., shelves or waves) or with no physical arrangement at all (e.g., numbers or ideas), we might expect stubbornly distributive predicates to more readily admit collective interpretations when the nouns that name these objects serve as subject. This reasoning leads to the prediction that the nouns which serve as subjects in a plural predication should also affect the likelihood of collective interpretation. In Expt. 2, we test whether this holds for a variety of predicates.

Beyond contextual predictability of collective properties, the pragmatic view of plural predication entails that other standard pragmatic pressures should come to bear on the reasoning processes of speakers and listeners, affecting the felicity and thereby likelihood of the collective (and distributive) interpretation. To whit, changing the speaker's epistemic state ought to influence the resulting disambiguation between distributive and collective interpretations. If a speaker lacks evidence that would verify a specific interpretation, and if a listener knows this about the speaker, then that unsupported interpretation should be particularly improbable. This intuition derives straightforwardly from the Maxim of Quality from Grice (1975): "Do not say that for which you lack adequate evidence." We test this prediction first in Expt. 1 with a reference task and then with a more sensitive measure in Expt. 3.

Before testing the predictions of our pragmatic account, we must establish a way of studying plural predication experimentally; that is, a way to unambiguously access distributive vs. collective interpretations. To that end, in Expt. 1, we show that the particles *each* and *together* reliably disambiguate plural predication.

2. Experiment 1: Validating the paraphrase methodology

One way to access interpretations of potentially ambiguous plural predications is to elicit ratings of unambiguous paraphrases. We construct these paraphrases using distributive *each* and collective *together*, as in (4).

[1] In his characterization of stubborn distributivity, Schwarzschild (2011) first assumes that gradable adjectives are predicates of events (e.g., Higginbotham & Schein, 1989). Thus, *heavy* predicates over HEAVY events, *big* over BIG events, etc. For Schwarzschild, stubbornly distributive predicates are those that require their events to have only single participants. The details are not directly relevant for our purposes; what is relevant is the lexical distinction Schwarzschild and others make between stubbornly distributive and complaisantly collective predicates (i.e., those predicates that allow collective construals).

[2] Similarly, truly collective predicates like *gather* or *meet* name properties that require a collective interpretation by virtue of their meaning (Schwarzschild, 1994).

[3] See Schein (2017, chap. 4.1.1) for a similar line of reasoning; Schein frames his observation in a Neo-Davidsonian event-semantic framework.

(4) The boxes were big/heavy/tall.
a. The boxes **each** were big/heavy/tall.
b. The boxes **together** were big/heavy/tall.

Although many theorists assume the success of these disambiguating particles (e.g., Schwarzschild, 1994), diligence is due: we must establish that these are in fact unambiguous paraphrases in the context of our experimental paradigm.[4] We do so by asking participants to choose between two collections of boxes as the referent of each definite description in (4), (4a), and (4b); the objects in one collection hold the property individually (but not collectively), while the other collection holds the property (but the individuals in it do not). We expect the two disambiguated sentences to lead to robust selection of the corresponding collection, while selections for the ambiguous sentence will indicate preferences for its interpretation. In addition, we begin to explore the role of context by manipulating the plausible knowledge of the speaker: whether he is likely to have access to the relevant information about each object (by, e.g., inspecting each box), or only about the collection (by, e.g., move the boxes all together).

2.1. Participants

We recruited 50 participants (21 female, 29 male; mean age: 31) with U.S. IP addresses through Amazon.com's Mechanical Turk crowd-sourcing service. Participants received $0.25 for their participation.

2.2. Design and methods

Participants were introduced to Cubert, a worker in a box factory.[5] To manipulate Cubert's access to knowledge about the boxes, between participants, the context scenario either had Cubert "move" or "inspect" the boxes that he received from a dispenser. In "move" scenarios, Cubert appeared with a cart positioned under the dispenser (Fig. 1, left); in "inspect" scenarios, there was no cart (Fig. 1, right). Participants were told that after moving/inspecting a shipment of boxes, Cubert told his friend Dot about the boxes he moved/inspected. Their task was to help Dot decide which boxes Cubert was referring to with his utterance. Participants chose between two sets of boxes, one that implied a collective interpretation of the utterance (five small boxes, Fig. 2, left) and one that implied a distributive interpretation (two large boxes which were together smaller than the five small boxes, Fig. 2, right).

Cubert used three predicates, *big*, *heavy*, and *tall*, in three sentence frames, or utterances: "bare" (4), "each" (4a), and "together" (4b). Participants saw each version of each predicate in a random order, completing nine trials. All 50 participants indicated that they were native speakers of English, so we included all of the data in the analyses reported below.

2.3. Predictions

If *each* and *together* are clear disambiguators for plural predication, we should find a split in the choice of referent depending on which of these words appears. Encountering distributive *each*, as in (4a), participants should choose the distributive referent (Fig. 2, right); with collective *together*, (4b), participants should choose the collective referent (Fig. 2, left).

For the potentially ambiguous bare utterances, stubbornly distributive *big* ought to always prefer distributive referents, while complaisantly collective *heavy* should be split. If predicates of size and shape are always stubbornly distributivity, *tall* should pattern with *big*. Given that size and height are visually assessable properties, the scenario manipulation should have no effect on Cubert's epistemic state, and therefore no effect on the resulting interpretations for bare *big* and *tall*. However, given that Cubert would need to lift individual boxes to assess their individual weight, we should find an effect of scenario for *heavy*: moving all of the boxes (as opposed to inspecting each box) would yield higher rates of collective referent choice; without lifting each box, Cubert would lack the knowledge needed to be justified in using a distributive interpretation.

2.4. Results

Results were coded in terms of whether participants chose the collective referent, and therefore accessed the collective interpretation for Cubert's statement to Dot. Fig. 3 displays the proportion of collective choices, with bootstrapped 95% confidence intervals drawn from 10,000 samples of the data (DiCiccio & Efron, 1996); a value of 1 indicates that participants chose the collective referent 100% of the time.

To evaluate the disambiguating potential of our paraphrases, we first look at responses to these paraphrases, excluding the bare utterance. Mean proportion of collective referent choice for collective paraphrases with *together* was 85% (i.e., near ceiling); for distributive paraphrases with *each* this proportion was 10% (i.e., near floor). We fit a mixed effects logistic regression model (Baayen, Davidson, & Bates, 2008) using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014) in *R*, predicting referent choice by UTTERANCE ("each," "together") and its interaction with PREDICATE (*big*, *heavy*, *tall*; *big* was coded as the reference level), as well as TRIAL order. The model included random intercepts for participants and scenarios (i.e., the maximal random effects structure supported by the data). The full regression model appears in Appendix A. The model finds a main effect of UTTERANCE ($\beta = 4.99$, $SE = 0.82$, $z = 6.08$, $p < 0.001$), confirming that collective *together* yielded greater rates of collective referent choice than distributive *each*; the interactions with PREDICATE were not significant (*heavy*: $\beta = -1.39$, $SE = 0.91$, $z = -1.52$, $p = 0.13$; *tall*: $\beta = -0.63$, $SE = 0.95$, $z = -0.66$, $p = 0.51$). The effect of TRIAL was also not significant ($\beta = 0.07$, $SE = 0.07$, $z = 1.00$, $p = 0.32$). Thus, the "together" utterances yielded collective referent choices, and the "each" utterances did not, for each of the three predicates.

To evaluate the behavior of the predicates in the absence of the disambiguating particles, together with the effect of our scenario manipulation, we next consider responses to the bare, potentially ambiguous utterance. We fit a mixed effects logistic regression model predicting referent choice by PREDICATE (*big*, *heavy*, *tall*), SCENARIO ("move," "inspect"), and TRIAL order. We dummy coded the PREDICATE predictor, with *big* as the reference level. The model included random intercepts and slopes for participants (grouped by TRIAL; the maximal random effects structure supported by the data). The full regression model appears in Appendix A. The model finds main effects of PREDICATE for both the *heavy* ($\beta = 3.05$, $SE = 1.23$, $z = 2.48$, $p < 0.05$) and the *tall* ($\beta = 3.63$, $SE = 1.29$, $z = 2.82$, $p < 0.01$) contrasts. Compared to *big*, the other predicates yielded greater rates of collective referent choice in "bare" utterances. While the effect of SCENARIO did not reach significance, we note the trend in Fig. 3 whereby "move" scenarios tended to have higher rates of collective referent choice for bare *heavy* utterances (34% "move" vs.

[4] Previous experimental work has investigated the disambiguating potential of *each* and so-called "post-VP" *together* (i.e., *the boxes were big together*). Syrett and Musolino (2013, 2016) found that these particles reliably disambiguate between distributive and collective interpretations for adults. However, we are aware of no empirical work on post-NP *together*, which we employ in our experimental materials.

[5] The full experiment is viewable online at http://cocolab.stanford.edu/experiments/collective/expt1/expt1.html.

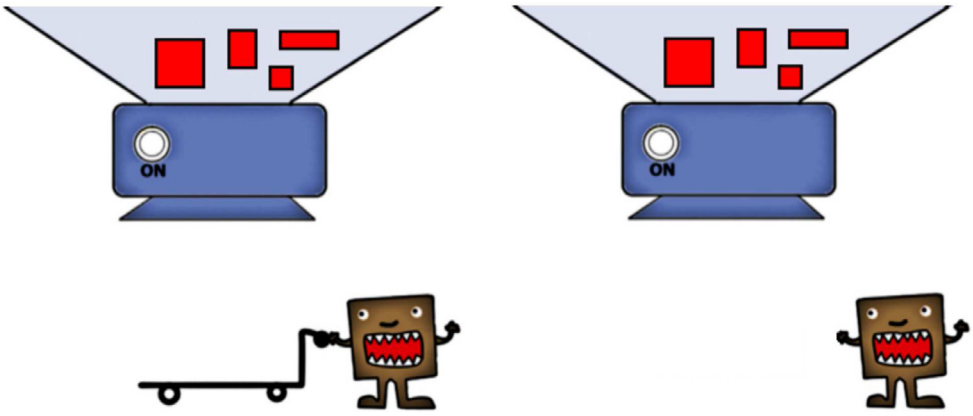

Fig. 1. Example contexts from Expt. 1. Left: "move" scenario with dolly. Right: "inspect" scenario without dolly.

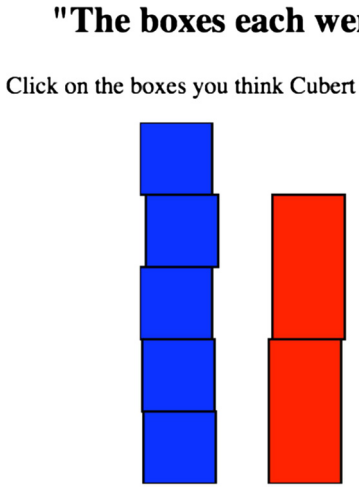

Fig. 2. Example trial from Expt. 1 with the predicate *big* in the "each" utterance.

14% "inspect"); a post hoc test of this difference approaches significance ($t = -1.70$, $df = 47.96$, $p = 0.096$).

2.5. Discussion

The disambiguating paraphrases behave as expected: *each* delivers a distributive interpretation and *together* delivers a collective one. We also find that the bare form of *big* patterns differently from *heavy* and *tall*—it is more distributive. This finding is expected with respect to *big* vs. *heavy* (the former should be stubbornly distributive), but surprising given the behavior of *tall*: both *tall* and *big* are predicates of size and shape. If naming properties of physical extent mediates the choice between distributive and collective interpretations of plural predications, we should find that *tall* patterns with *big*. Equally damning is the comparison between *tall* and *heavy*: if anything, *tall* more consistently delivers collective interpretations. What allows *tall* to readily receive collective interpretations—while precluding them for *big*—in our experimental context?

As discussed above, one candidate factor is the contextual predictability of the collective property: how easy it is for speakers and listeners to arrive at the same collective property for a given set of objects. We have seen that collective weight is a stable property of sets, which stands to explain the complaisance of *heavy* toward collective interpretations. In the context of the current experiment, collective height is also predictable: boxes always appeared stacked one on top of the other, yielding stable total set heights. Collective size, however, could vary depending on the strategy used to evaluate it (height? width? area? volume?) as well as on the arrangement of the objects said to hold it, thus accounting for the lack of collective interpretations for bare *big*. In Expt. 3, we use the paraphrase methodology to directly investigate the role of contextual predictability in plural predication.

The results of the current experiment also leave open the possibility of the influence of speaker knowledge in plural predication: based solely on whether the context scenario had Cubert move vs. inspect the boxes he received, we see a trend whereby bare *heavy* received more or less collective interpretations, respectively. The absence of the effect for *big* and *tall* suggests an explanation in terms of the epistemic state of the speaker: in "move" scenarios, the speaker is less likely to have access to individual box weights (having plausibly moved them *en masse*); the listener is therefore less likely to assume that the speaker intended a distributive interpretation for which he lacks evidence. We attempt to confirm this effect with a similar scenario manipulation in Expt. 3.

Before manipulating contextual predictability and speaker knowledge directly, we extend the paraphrase methodology in

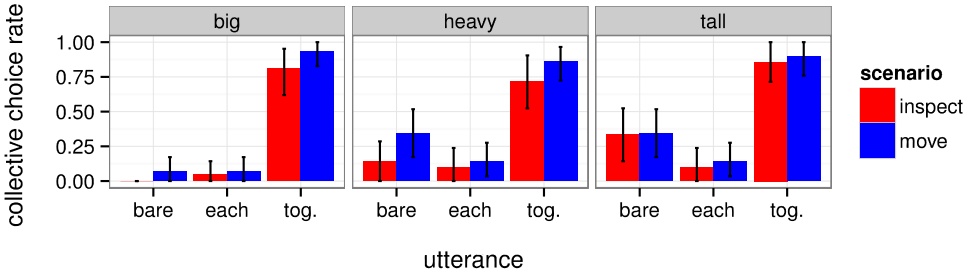

Fig. 3. Proportion of collective choices in the results of Expt. 1.

Expt. 2 to naturally-occurring examples of plural predication, investigating the role of the local linguistic context.

3. Experiment 2: Naturally-occurring examples

In Expt. 1, we used a reference task to validate disambiguating paraphrases as a probe of interpretations of plural predications. We next use the paraphrase methodology to evaluate natural plural predications gathered from corpora. We aim to both extend our understanding of distributive vs. collective preferences to predicates beyond *big*/*tall*/*heavy*, and explore the effect of linguistic context: how the subject noun affects this preference. We thereby test the null hypothesis of a simple lexical distinction between stubbornly distributive and complaisantly collective predicates: if stubbornly distributive predicates are truly stubborn, we should find a clear split in behavior between them and the other predicates, and we should find no effect of the subject noun on their resistance to collective interpretations.

3.1. Experiment 2a: Frequent plural predications

We begin with naturally-occurring examples of plural predications extracted from the British National Corpus.[6]

3.1.1. Participants

We recruited 90 participants (46 female, 44 male; mean age: 36) with U.S. IP addresses through Amazon.com's Mechanical Turk. Participants received $0.30 for their participation.

3.1.2. Design and methods

In a search of the British National Corpus, we identified the 40 most frequent subject-predicate combinations in the plural predication frame in (5).

(5) The NOUNS were ADJECTIVE.

Each participant was presented with 30 sentences; 15 of these sentences were drawn randomly from the original set of the 40 most frequent instances. The other sentences were constructed by randomly pairing subjects and predicates from the original set of 40 to create 15 novel sentences for each participant.

Each sentence was presented as an utterance by a speaker, and participants were tasked with determining what the speaker meant.[7] Participants rated two possible paraphrases on a sliding scale. The paraphrases were distributive (with "each", (6a)) and collective (with "together", (6b)).

(6) a. The NOUNS each were ADJECTIVE.
b. The NOUNS together were ADJECTIVE.

Given our interest in rates of collective interpretations, we are primarily concerned with collective paraphrase endorsement rates; we expect distributive paraphrase endorsement rates to give redundant information. We included both distributive and collective paraphrases to highlight the potential ambiguity for participants. Scale endpoints corresponded to whether the paraphrase was "definitely" or "definitely not" what the speaker intended by his or her utterance. Before rating its paraphrases, participants judged whether the utterance made sense (choosing between "Yes" and "No"). We analyzed data from the 85 participants who indicated that their native language was English.

3.1.3. Results

Of the original 40 most frequent sentences, participants indicated that they "made sense" 95% of the time.[8] To avoid possible confusion that nonsensical interpretations might introduce, in the analyses reported below we only look at responses to utterances that participants said made sense.

We begin with a look at the range of collective endorsement ratings.[9] Fig. 4 plots average collective endorsement ratings for each of the 40 most frequent sentences with bootstrapped 95% confidence intervals. Descriptively, we see a wide range of ratings, spanning from approximately 25% to 85% endorsement; participants appear to be comfortable using the full scale. An important observation about the distribution of ratings in Fig. 4 is its relatively smooth gradient: we fail to find clear groupings of predicates, such that they behave in one way or another with respect to collective interpretations. Stubbornly distributive *small* indeed occurs toward the low end of collective endorsement ratings, but not markedly so.

There are 8 predicates that occur with multiple nouns in our small corpus. We plot their collective endorsement ratings in Fig. 5, shifting to violin plots to give a better picture of the ratings distribution. The plots also include mean collective endorsement ratings and bootstrapped 95% confidence intervals. For at least 3 out of these 8 predicates, rates of collective endorsement differ significantly by subject noun (*bright*, *closed*, *small*; see Appendix B for details). We draw the reader's attention to the case of stubbornly distributive *small* (bottom right of Fig. 5). If stubborn distributivity were a binary, predicate-level phenomenon hard-coded into the semantics, we should find no effect of the subject noun in the determination of *small*'s distributive vs. collective interpretations: stubbornly distributive predicates, regardless of their subject, should always be maximally distributive. But this is not the case for *small* in Fig. 5. A linear mixed effects model predicting collective endorsement rates by subject noun for *small*, with random by-participant intercepts, finds that sentences with the subject *numbers* received greater rates of collective endorsement than sentences with *children* ($\beta = -0.15$, $SE = 0.07$, $t = -2.24$, $p < 0.05$), *rooms* ($\beta = -0.19$, $SE = 0.07$, $t = -2.73$, $p < 0.01$), or *classes* ($\beta = -0.27$, $SE = 0.07$, $t = -3.87$, $p < 0.01$).[10] Curiously, *numbers* names objects that do not instantiate physically, which plausibly increases the predictability of their collective size (i.e., their summation) and thus increases rates of collective interpretations.

This analysis relies on the chance occurrence of multiple subject nouns in our small corpus to test the effects of local linguistic context. In Expt. 2b, we follow up on this result by systematically testing the effect of subject nouns on a narrower set of predicates.

3.2. Experiment 2b: big, heavy, tall

We then narrowed our sights to the predicates tested in Expt. 1: *big*, *heavy*, and *tall*. These predicates are of particular interest because they come from both supposed lexical categories (stubbornly distributive and complaisantly collective) and the physical

[6] Sentences were extracted from the British National Corpus Online service, managed by Oxford University Computing Services on behalf of the BNC Consortium. All rights in the texts cited are reserved.

[7] The full experiment is viewable online at http://cocolab.stanford.edu/experiments/collective/expt2a/expt2a.html.

[8] Of the novel sentences, participants indicated that they "made sense" 67% of the time.

[9] A look at distributive endorsement ratings, or normalized collective ratings (i.e., the difference between collective and distributive ratings) reveals similar patterns of results, owing to the high negative correlation between distributive and collective endorsement ratings ($r = -0.78$).

[10] We find the same pattern of results if instead we compare each subject noun with the grand mean (i.e., through effect, or deviation, coding of our variables): *numbers* deviates significantly, while the other nouns do not.

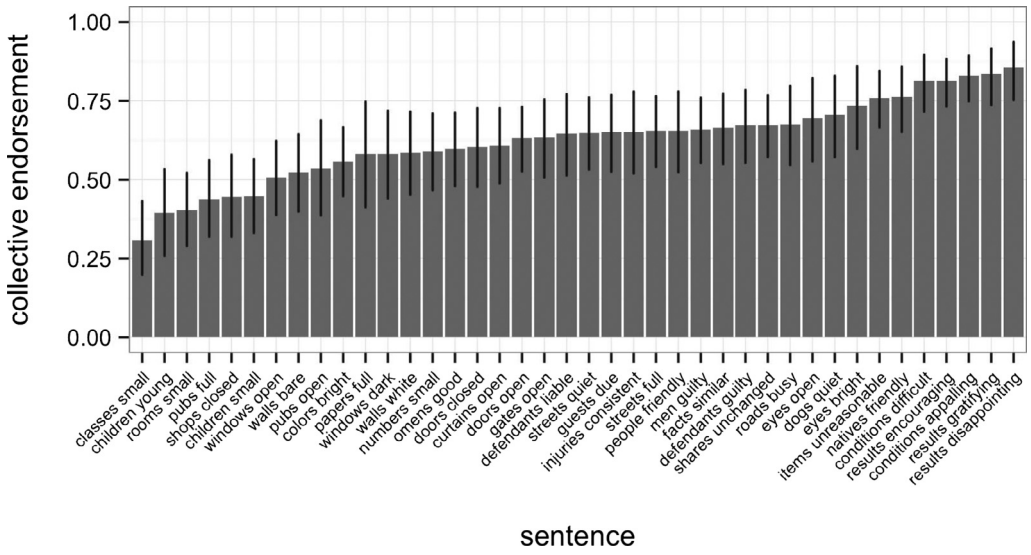

Fig. 4. Collective paraphrase endorsement rates for different noun-predicate combinations in Expt. 2a (i.e., for different sentences).

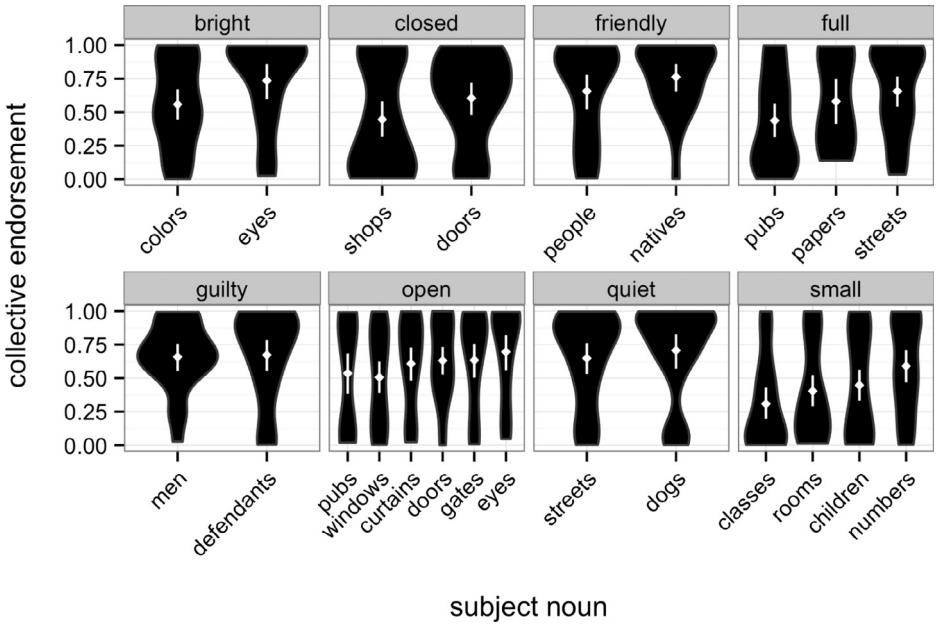

Fig. 5. Collective paraphrase endorsement ratings for different subject nouns grouped by predicate (Expt. 2a).

properties they name lend themselves to direct manipulation. Due to the limited number of instances of plural predication with these predicates in the British National Corpus, we use instead the much larger Google Books Corpus (Davies, 2011) to look at their most frequent subject nouns.

3.2.1. Participants

We recruited 30 participants (25 female, 5 male; mean age: 38) with U.S. IP addresses through Amazon.com's Mechanical Turk. Participants received $0.30 for their participation.

3.2.2. Design and methods

The design and methods for this experiment were identical to those of Expt. 2a, with the exception that our naturally-occurring examples feature the predicates *big*, *heavy*, and *tall*, and derive from the Google Books Corpus. We identified the five most frequent subject nouns in the plural predication frame (i.e., *the* NOUNS *were* ADJECTIVE) for each of the predicates *big*, *heavy*, and *tall*.[11] The subject-predicate pairs appear in (7).

(7) *Noun-predicate pairings*:

big:	boys, children, houses, rooms, waves
heavy:	bags, lids, loads, men, trees
tall:	buildings, offspring, plants, trees, windows

[11] For *heavy*, many of the most frequent subject nouns resulted in a non-canonical, abstract interpretation. For example, *heavy rains* or *heavy casualties*. We excluded these cases from the current experiment. In a separate experiment, we included just these abstract instances of plural predication with *heavy* and found no effect of the subject noun on interpretation.

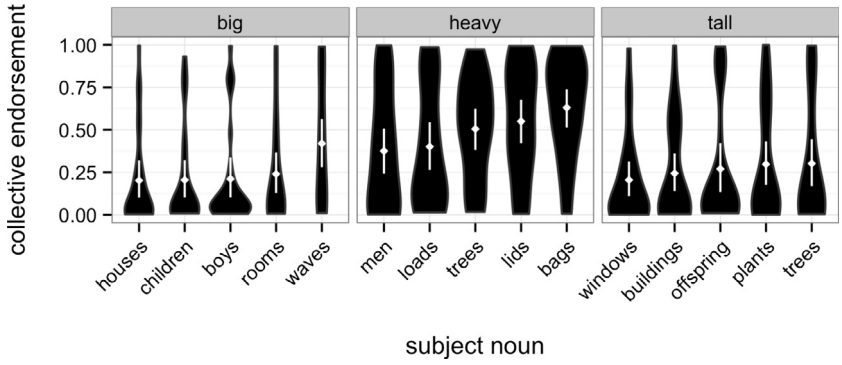

Fig. 6. Collective paraphrase endorsement ratings for different nouns serving as the subject to the predicates *big*, *heavy*, and *tall* (Expt. 2b).

As with Expt. 2a, participants judged whether the resulting sentences made sense and rated disambiguating paraphrases of each.[12] We analyzed data from the 26 participants who indicated that their native language was English.

3.2.3. Results

Of the original 15 most frequent sentences, participants indicated that they made sense 98% of the time.[13] We again look only at responses to the original most frequent utterances that participants indicated made sense to them.

Fig. 6 plots collective endorsement ratings for each of the three predicates grouped by their subjects.[14] To compare the predicates with each other, we fit a linear mixed effects model predicting collective endorsement by predicate, with random intercepts for participants and nouns, and random slopes for participants. The predicate predictor was dummy coded with *heavy* as the reference level. There were main effects for both contrasts, such that *big* ($\beta = -0.22$, $SE = 0.07$, $t = -3.24$, $p < 0.01$) and *tall* ($\beta = -0.20$, $SE = 0.06$, $t = -3.20$, $p < 0.01$) led to lower rates of collective endorsement than *heavy*.

We might take the split in behavior between *heavy* on the one hand and *big* and *tall* on the other as evidence for a categorical divide between complaisantly collective and stubbornly distributive predicates. However, comparing the predicates at the level of individual subject nouns suggests otherwise: the subject noun with the highest rates of collective endorsement for *big*, *waves*, yields rates of collective endorsement that do not differ significantly from *heavy*'s least collective subject noun, *men* (42% *big waves* vs. 38% *heavy men*; $\beta = -0.04$, $SE = 0.09$, $t = -0.51$, $p = 0.62$). The same holds of the comparison between *tall*'s most collective subject noun, *trees*, and *heavy men* (30% *tall trees*; $\beta = -0.07$, $SE = 0.08$, $t = -0.87$, $p = 0.39$).

Next, we compare the behavior of the subject nouns within each predicate. Starting with *tall*, we fit a linear mixed effects model predicting collective endorsement by subject noun, with random intercepts for participants; full regression models appear in Appendix B. The subject noun predictor was dummy coded with *trees* as the reference level (i.e., the subject with the highest average collective endorsement rating). Compared to *trees*, we do not find an effect of the contrast with *plants* ($\beta = 0.00$, $SE = 0.06$, $t = -0.07$, $p = 0.95$), *offspring* ($\beta = -0.03$, $SE = 0.06$, $t = -0.55$, $p = 0.59$), *buildings* ($\beta = -0.06$, $SE = 0.06$, $t = -1.00$, $p = 0.32$), or *windows* ($\beta = -0.09$, $SE = 0.06$, $t = -1.52$, $p = 0.13$). In other words, *tall*'s subjects behave similarly in their general resistance to collective interpretation.

Turning to *big*, we performed a similar analysis, this time dummy coding *waves* (the subject with the greatest average collective endorsement rating) as the reference level. There were significant effects of each contrast: *rooms* ($\beta = -0.18$, $SE = 0.06$, $t = -2.97$, $p < 0.01$), *boys* ($\beta = -0.22$, $SE = 0.06$, $t = -3.53$, $p < 0.01$), *children* ($\beta = -0.21$, $SE = 0.06$, $t = -3.54$, $p < 0.01$), and *houses* ($\beta = -0.23$, $SE = 0.06$, $t = -3.68$, $p < 0.01$); *waves* stands out with its relatively high levels of collective endorsement for *big*.

Finally, *heavy*; here, *bags* received the highest collective endorsement ratings, so we coded it as the reference level. The contrast with *lids* was not significant ($\beta = -0.08$, $SE = 0.08$, $t = -1.04$, $p = 0.30$), the contrast with *trees* was marginally significant ($\beta = -0.15$, $SE = 0.08$, $t = -1.93$, $p = 0.06$), and the contrasts with *loads* ($\beta = -0.24$, $SE = 0.08$, $t = -3.15$, $p < 0.01$) and *men* ($\beta = -0.25$, $SE = 0.08$, $t = -3.56$, $p < 0.01$) were significant. In other words, we find similarly high ratings for *bags* and *lids*, intermediate ratings for *trees*, and similarly low ratings for *loads* and *men*.

3.3. Discussion

We have extended our paraphrase methodology beyond a simple reference task, using naturally-occurring examples of plural predications. The pattern of ratings sheds some light on the status of stubborn distributivity. Perhaps most striking is the absence of clear groupings of predicates in the results of Expt. 2a: one would be hard-pressed to read stubborn distributivity (or complaisant collectivity) off the gradient of ratings in Fig. 4. Still, one might wonder about the precise meanings of the collective interpretations of some of our sentences, for example "the boys together were smiling" or "the natives together were friendly." To obviate the murky evaluation criteria of collective action statements like these, we have chosen to consider in detail a specific class of predicates with (relatively) objective evaluation strategies: gradable adjectives like *heavy* or *big*.[15] The measurement inherent to the semantics of these predicates delivers scalar interpretations that are more amenable to quantitative study. By examining intuitions about the use of these predicates, together with the predication context that delivers them, we stand to better understand the mechanism of plural predication: when collective interpretations are

[12] The full experiment is viewable online at http://cocolab.stanford.edu/experiments/collective/expt2b/expt2b.html.

[13] Of the novel sentences, participants indicated that they made sense 87% of the time.

[14] As in Expt. 2a, distributive endorsement ratings provide redundant information, owing to their high negative correlation with the collective ratings ($r = -.98$).

[15] This is not to say that we know nothing about collective action statements. For example, Margaret Gilbert's Plural Subject Theory makes great gains on this ground in the domain of social philosophy (for a recent overview, see Gilbert, 2013).

appropriate, what they communicate, and what mediates the choice between distributive and collective interpretations in the first place.

We therefore narrowed our focus to the predicates *big*, *heavy*, and *tall* in Expt. 2b, finding that complaisantly collective *heavy* indeed does yield greater rates of collective paraphrase endorsement, and that stubbornly distributive *big* and *tall* pattern together in their resistance to collective interpretations. However, once we consider the effect of local linguistic context by comparing responses to specific subject nouns, the division between the two classes of predicates disappears. *Big* and *tall* are no longer so stubborn with the appropriate context (cf. the relatively high rate of collective interpretations participants demonstrated for *tall* in Expt. 1).

By comparing specific subject nouns within predicates, we begin to find evidence for the context-dependent nature of these predications. With *big*, four out of the five subject nouns strongly resisted collective interpretations. However, with *waves*, participants much more readily endorsed collective paraphrases (as was the case with *small numbers* in Expt. 2a). Several interpretations of this finding are possible; one is that limiting the noise of a collective interpretation—by reducing the variability in the physical instantiation of the subject—increases the rates of that interpretation. We follow up on this hypothesis with a direct manipulation of predictability in Expt. 3.

With *heavy*, we also found that different subject nouns yielded different rates of collective interpretations. At first blush, these differences appear to track the probability that a speaker would have access to the information necessary to verify a distributive interpretation. Lids and bags are probably picked up together, certainly more so than loads or men. If speakers experience only the collective weight of a plurality, not the individual weights of the members, they lack knowledge which would license the distributive meaning. If listeners are aware of this, then the alternative, collective reading becomes more likely for collections (such as lids) that are likely to be lifted together. As with contextual predictability, we follow up on this result with a direct manipulation of speaker access to knowledge in Expt. 3.

4. Experiment 3: Manipulating context

Physical properties of collections will tend to be more predictable when the physical arrangement is more predictable. According to our hypothesis, greater predictability should lead to more collective interpretations. For instance, if collections of certain boxes are always stacked neatly, the collective height of a stack can be stably predicted (as the sum of individual heights); if these boxes come in jumbled-up piles the collective height is less predictable. In Expt. 3, we manipulate expectations about the arrangement of boxes, using the paraphrase endorsement methodology to test the effect on the interpretation of plural predications. We expect to see an effect on *tall*, whose collective property is the direct target of our manipulation; we also expect to see a more moderate effect on *big* to the extent that *big* can refer to height; we expect no effect on *heavy*, which is stable regardless of arrangement.

In Expt. 3, we also return to the effect of speaker's knowledge. Recall the prediction of our pragmatic hypothesis, that more collective interpretations should attain when the speaker lacks knowledge about individual object properties (but has knowledge about collective properties). We conjectured that the effect of *heavy*'s subject nouns in Expt. 2 was due to knowledge in this way: for smaller objects the likelihood of interacting with them individually is higher, making the likelihood of individual weight knowledge higher. In Expt. 1, we attempted to manipulate the likelihood that the speaker had interacted with the boxes individually, via the "inspect" scenario, versus only collectively, in the "move" scenario. We found a trend in the predicted direction that did not reach significance. It is possible that this effect was real, but small because of a bias for the distributive referent (leading to a floor effect). In Expt. 3, we repeat the move/inspect manipulation using the pure paraphrase endorsement dependent measure of Expt. 2, which may have greater sensitivity than a binary choice between possible referents.

4.1. Participants

We recruited 80 participants (26 female, 54 male; mean age: 32) with U.S. IP addresses through Amazon's Mechanical Turk. Participants received $0.25 for their participation.

4.2. Design and methods

The experimental context was similar to that of Expt. 1. Participants were first introduced to an agent, Cubert, who works in a factory with boxes. Cubert receives boxes from a dispenser in the ceiling. Participants began by observing the dispenser in action.[16]

Between subjects, we manipulated the SCENARIO story ("move" vs. "inspect"): Cubert's job was either to move shipments of boxes, or to inspect them. As in Expt. 1, in "move" scenarios, Cubert appeared with a dolly with which to carry the boxes. In "inspect" scenarios, Cubert appeared without a dolly.

Also between subjects, we manipulated the variability of the CONTEXT ("regular" vs. "random"). In regular contexts, the dispenser consistently stacked boxes on top of each other (Fig. 7, left). In random contexts, boxes were dispensed without any consistent physical arrangement (Fig. 7, right). Participants saw a series of four context priming scenarios; in each scenario, they saw a different set of boxes dispensed.

After observing the dispenser in action via the context scenarios, participants were then introduced to a second agent, Dot, whom Cubert told about some boxes he moved/inspected. Participants were tasked with helping Dot understand what Cubert meant by his utterance. To do so, they saw an utterance and rated potential paraphrases on a sliding scale (as in Expt. 2). The paraphrases were distributive (with "each") and collective (with "together"). Scale endpoints corresponded to whether the paraphrase was "definitely" or "definitely not" what Cubert meant by his utterance (Fig. 8). (Note that, unlike Expt. 1, there was no explicit referent visible.) Participants completed three trials in a random order with three different predicates: *big*, *heavy*, and *tall*. For the analyses reported below, we included data from the 77 participants who indicated that their native language was English.

4.3. Results

Fig. 9 displays violin plots of the raw paraphrase endorsement ratings, with mean endorsement ratings and bootstrapped 95% confidence intervals. For the purpose of analysis, we analyze responses to the collective paraphrases (i.e., the red bars in Fig. 9); higher rates of endorsement for the collective paraphrase signal greater rates of collective interpretation. First note the qualitative effects of our two manipulations: collective interpretations do appear to be greater for *tall* and *big* in the "regular" conditions, and greater for *heavy* in the "move" conditions.

For each predicate, we fit a linear regression model predicting collective endorsement ratings by CONTEXT ("random" vs. "regular"), SCENARIO ("inspect" vs. "move"), and TRIAL order; the model also

[16] The full experiment is viewable online at http://cocolab.stanford.edu/experiments/collective/expt3/expt3.html.

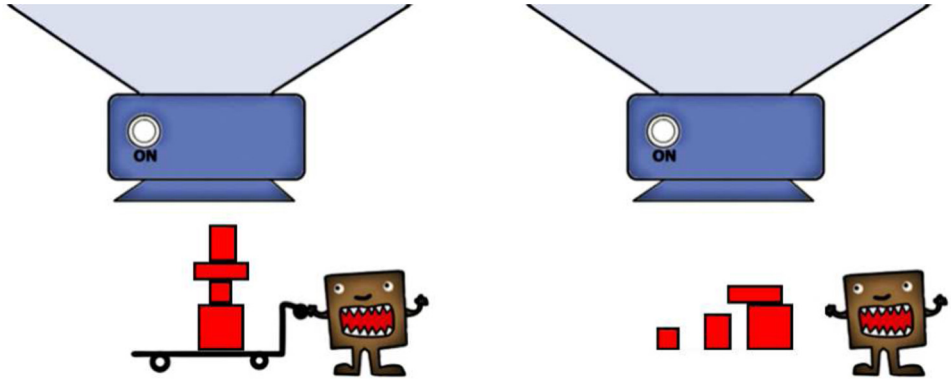

Fig. 7. Example context scenarios from Expt. 3. Left: the outcome of a "regular," "move" context scenario. Right: the outcome of a "random," "inspect" context scenario.

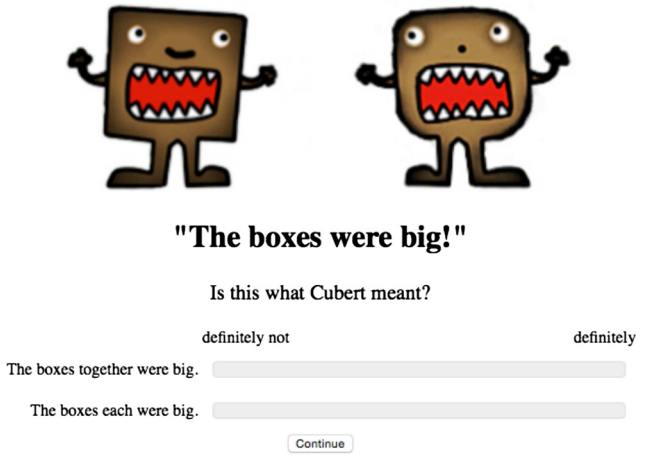

Fig. 8. Example "big" trial.

included the interaction between CONTEXT and SCENARIO.[17] The full regression models appear in Appendix C. As is apparent in Fig. 9, there was a main effect of CONTEXT for both *big* (β = 0.18, SE = 0.07, t = 2.52, $p < 0.05$) and *tall* (β = 0.30, SE = 0.07, t = 4.58, $p < 0.01$), such that regular contexts had more collective interpretations than did random contexts for these predicates. The effect of CONTEXT was not significant for *heavy* (β = 0.11, SE = 0.07, t = 1.43, $p = 0.16$), but the effect of SCENARIO was (β = 0.19, SE = 0.08, t = 2.54, $p < 0.05$): *heavy* received more collective interpretations in "move" scenarios. No other effects reached significance.

4.4. Discussion

As we would expect if contextual predictability of collective properties influences the viability of collective interpretations, regular contexts had higher ratings for collective paraphrases (and lower ratings for distributive ones). The predicate *tall* was most affected by our contextual manipulation, which directly targeted the height dimension, stacking boxes on top of each other. But *big*, also claimed to be stubbornly distributive, is similarly affected: regular contexts yield more collective interpretations.

For the predicate *heavy*, our contextual predictability manipulation had no measurable effect, presumably because collective weight is already maximally predictable. However, *heavy* was sensitive to the "move" vs. "inspect" scenario manipulation: as was the trend in the results of Expt. 1, "move" scenarios yielded much greater collective endorsement ratings for *heavy*. This manipulation privileges collective interpretations by limiting speakers' access to the information they would need to verify a distributive interpretation; Cubert lacked access to individual box weights in "move" scenarios. The predicates *big* and *tall* were not affected by the speaker knowledge scenario manipulation, owing to the fact that they name properties that are visually accessible, and thus visually assessable.

An alternative interpretation of the contextual predictability effect is that in regular, stacked contexts, participants reanalyzed the plural definite description *the boxes* as referring to a single individual (e.g., to a single pile, and not to a set of boxes, perhaps by imagining a silent collectivizing noun); if *the boxes* referred to a single individual, there never was any collective predication. Such a representational coercion story raises serious problems for the semantics of plural definite descriptions (for discussion, see Link, 1983; Landman, 1989a; Schwarzschild, 1996; Link, 1998). More importantly it is inconsistent with the *lack* of effect for *heavy*: if the pile of boxes was simply represented as an individual when stacked, we would expect greater collective interpretation rates for all predicates in regular contexts.

Taken together, our results demonstrate the central role of context in plural predication: as collective properties become more predictable, collective interpretations become more likely; and without epistemic support, distributive interpretations become less likely. These findings support a pragmatic account of stubborn distributivity according to which ostensibly stubbornly distributive predicates are such because the properties they name are unpredictable, or unstable in most contexts. In the next experiment we probe the generalizability of our findings.

5. Experiment 4: Generalizing our findings

In Expt. 3, we looked at the role of contextual predictability in interpretation choice for three adjectives of English: *big*, *heavy*, and *tall*. One might worry that the observed success of our contextual manipulation is an artifact of the small set of adjectives we tested, and might not generalize to a broader set of adjectives. Therefore, to test the generalizability of the findings from Expt. 3, we used the design on a much broader set of 25 dimensional adjectives.

[17] All reported results hold if instead we look at normalized collective paraphrase endorsement ratings (i.e., the difference between collective and distributive endorsements for each trial), given the strong negative correlation between the two endorsement ratings ($r = -0.62$). The results also hold if we analyze the results from all three predicates in a single mixed effects regression model; we present the three separate models here for ease of interpretation.

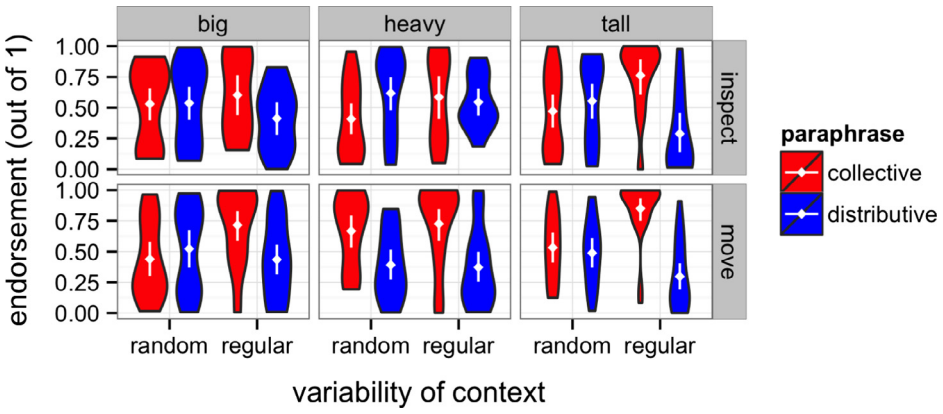

Fig. 9. Paraphrase endorsement ratings from Expt. 3.

5.1. Participants

We recruited 54 participants (21 female, 32 male; mean age: 35) with U.S. IP addresses through Amazon's Mechanical Turk.[18] Participants received $0.50 for their participation.

5.2. Design and methods

The experimental design was near-identical to that of Expt. 3. Two aspects of the design differed: there was no "move" vs. "inspect" scenario manipulation, and participants provided paraphrase endorsements for 25 rather than three predicates. At the outset of the experiment, participants were introduced to Cubert, who was tasked with "moving" boxes that he received from a dispenser in the ceiling (there was no "inspect" condition). Between subjects, we manipulated the variability of the CONTEXT ("regular" vs. "random") by having the dispenser either stack boxes on top of each other or dispense them in a random physical arrangement. Participants saw a series of four context priming scenarios to demonstrate the behavior of the dispenser; we used the same context priming scenarios from Expt. 3.

After observing the dispenser, participants were instructed to help Cubert's friend, Dot, understand Cubert's statements about the boxes he moved. To do so, they saw an utterance and rated potential paraphrases on a sliding scale; the paraphrases were either distributive (with "each") or collective (with "together"). Participants provided paraphrase endorsements for each of the adjectives in Table 1. We arrived at the set of adjectives in Table 1 by extracting every unique adjective appearing in an "A A N" NP from the Penn Treebank subset of the Switchboard corpus; there were 350 unique adjectives. Then, adjectives were independently coded according to semantic classes from the literature (e.g., Dixon, 1982). Among this set there were 24 unique dimensional adjectives; to those 24 we added the weight adjective *heavy*.

Participants completed 25 trials in a random order, one trial for each adjective. For the analyses reported below, we included data from the 51 participants who indicated that their native language was English.

5.3. Results

Fig. 10 displays the paraphrase endorsement ratings with means and bootstrapped 95% confidence intervals; higher paraphrase endorsement ratings indicate greater rates of the relevant interpretation. We grouped data by the dimension modified, together with adjective polarity.

Table 1
Adjectives, modified dimensions, and polarities tested in Expt. 4.

ADJECTIVE	DIMENSION	POLARITY
Full	Capacity	Positive
Deep	Depth	Positive
Flat	Height	Negative
Low	Height	Negative
Short	Height	Negative
High	Height	Positive
Tall	Height	Positive
Lengthy	Length	Positive
Long	Length	Positive
Little	Size	Negative
Mini	Size	Negative
Slight	Size	Negative
Small	Size	Negative
Tiny	Size	Negative
Big	Size	Positive
Huge	Size	Positive
Humongous	Size	Positive
Large	Size	Positive
Heavy	Weight	Positive
Narrow	Width	Negative
Skinny	Width	Negative
Thin	Width	Negative
Fat	Width	Positive
Thick	Width	Positive
Wide	Width	Positive

For each dimension, we fit a linear regression model predicting collective endorsement ratings by CONTEXT ("random" vs. "regular") and TRIAL order; for those dimensions with both positive and negative adjectives (i.e., the height, size, and width dimensions), models additionally included the POLARITY predictor and the interaction between CONTEXT and POLARITY.[19] The full regression models appear in Appendix D.

We begin with a look at the dimensions with only positive-polarity instances. For **capacity** adjectives, the model found a marginally significant effect of CONTEXT (β = 0.16, SE = 0.08, t = 1.95, $p < 0.06$), such that regular contexts tended to have more collective interpretations than did random contexts. However, repeating

[18] One participant chose not to provide a gender.

[19] All reported results hold if instead we look at normalized collective paraphrase endorsement ratings (i.e., the difference between collective and distributive endorsements for each trial), given the strong negative correlation between the two endorsement ratings ($r = -0.49$). The single exception is the marginally significant effect of CONTEXT for the capacity adjective *full*, which holds marginally for the collective endorsement ratings but not for the normalized ratings. We return to this point below.

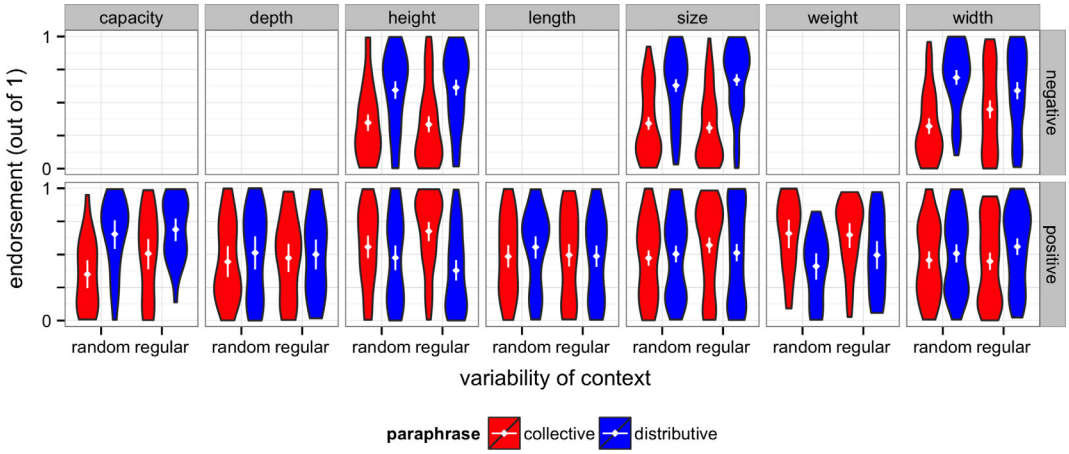

Fig. 10. Paraphrase endorsement ratings from Expt. 4.

the analysis for the normalized endorsement ratings, we fail to find any effect of CONTEXT ($\beta = -0.14$, $SE = 0.13$, $t = -1.04$, $p < 0.31$). There was also no significant effect of CONTEXT for **depth** ($\beta = 0.07$, $SE = 0.09$ $t = 0.85$, $p < 0.40$), **length** ($\beta = 0.01$, $SE = 0.06$, $t = 0.15$, $p < 0.89$), or **weight** ($\beta = -0.04$, $SE = 0.07$, $t = -0.51$, $p < 0.62$) adjectives.

We turn next to those dimensions with both negative- and positive-polarity instances. For **height** adjectives, the model found a main effect of polarity ($\beta = 0.28$, $SE = 0.04$, $t = 7.82$, $p < 0.01$), such that positive-polarity height adjectives received more collective interpretations than negative-polarity height adjectives. The model also found a marginally significant interaction between CONTEXT and POLARITY ($\beta = 0.13$, $SE = 0.07$, $t = 1.82$, $p < 0.07$); we found the same marginally significant interaction in an analysis of the normalized endorsement ratings ($\beta = -0.25$, $SE = 0.13$, $t = -1.96$, $p < 0.06$). A post hoc analysis revealed that this interaction was driven by a significant effect of CONTEXT within the positive-polarity height adjectives ($\beta = 0.12$, $SE = 0.06$, $t = 2.11$, $p < 0.05$): regular contexts yielded more collective interpretations. For **size** adjectives, the model found a main effect of polarity ($\beta = 0.20$, $SE = 0.03$, $t = 7.60$, $p < 0.01$), such that positive-polarity size adjectives received more collective interpretations than negative-polarity size adjectives. The model also found a significant interaction between CONTEXT and POLARITY ($\beta = 0.13$, $SE = 0.05$, $t = 1.82$, $p < 0.05$). A post hoc analysis revealed that this interaction was driven by a significant effect of CONTEXT within the positive-polarity size adjectives ($\beta = 0.10$, $SE = 0.04$, $t = 2.21$, $p < 0.05$): regular contexts yielded more collective interpretations. For **width** adjectives, the model found a significant interaction between CONTEXT and POLARITY ($\beta = -0.14$, $SE = 0.07$, $t = -2.02$, $p < 0.05$). A post hoc analysis revealed that this interaction was driven by a significant effect of CONTEXT within the *negative*-polarity width adjectives ($\beta = 0.13$, $SE = 0.05$, $t = 2.66$, $p < 0.01$): regular contexts yielded more collective interpretations.

5.4. Discussion

We have replicated the findings from Expt. 3 with a broader set of adjectives: *tall* and the three other positive-polarity height adjectives are affected by the contextual manipulation such that more predictable, "regular" contexts yield greater rates of collective interpretations; *big* and the four other positive-polarity size adjectives show the same effect of context. Moreover, the complaisantly collective weight adjective *heavy* continues to be unaffected by our contextual manipulation, which does little to increase the contextual predictability of the already-predictable collective weight. This replication serves double duty, first as a sanity check for comparing the results of Expt. 3 with the current experiment, and second as confirmation that the context effect observed in Expt. 3 generalizes beyond the specific lexical items that we tested.

In addition to replicating the effect of contextual predictability with a broader set of adjectives, we tested four more physical dimensions, as well as adjective polarity. Among those additional dimensions with only positive-polarity instances (i.e., capacity, depth, and length adjectives), none showed a reliable effect of our contextual manipulation. Thus, stacking boxes on top of each other failed to deliver greater rates of collective interpretations for adjectives modifying box capacity (i.e., *full*), depth (i.e., *deep*), or length (i.e., *long*, *lengthy*).[20] From this it would appear that generally regularizing physical arrangements is insufficient to deliver collective interpretations; rather, the dimension that the interpretation would target must be regularized. Given that stacking boxes plausibly regularizes collective height, size (to the extent that height factors into size), and width, it stands to reason that only these dimensions showed effects of our contextual manipulation. However, there is more to this story, because these dimensions showed sensitivity to our contextual manipulation according to their polarity.

Within the dimensions that showed clear effects of our contextual manipulation, positive—but not negative—height and size adjectives, and negative—but not positive—width adjectives had increased collective interpretations with increased contextual predictability. The strong influence of polarity highlights an important aspect of the pragmatics of utterance disambiguation that we have so far ignored: interpretations that are unlikely to be true are unlikely to be chosen—cooperative speakers do not intentionally utter false statements (Grice, 1975). Take the case of negative-polarity height adjectives like *short* or *flat*. It seems unlikely that Cubert would intend to communicate that a stack of boxes taller than him is collectively short when the distributive alternative is available, namely that each box is short. Stacking boxes may remove barriers to the collective interpretation by increasing the predictability of the collective property, but it creates its own barriers by highlighting that the collective property is unlikely to satisfy the truth conditions of the predicate.

[20] One might worry that *lengthy* in fact modifies a temporal dimension (e.g., *the conversations were lengthy*) that cannot felicitously be attributed to boxes. Still, all of the reported results hold if instead we analyze *long* and *lengthy* separately: neither is affected by our contextual manipulation.

It could be the case that negative-polarity dimensional adjectives are truly stubbornly distributive; the lack of a contextual effect for negative height and size adjectives would follow from this peculiarity of their lexical semantics. But then we should find the same pattern for the width adjectives, when in fact we find exactly the opposite pattern: an effect of context for the negative—but not positive—width adjectives like *narrow*. Again, the most straightforward interpretation of this pattern relies on the avoidance of false interpretations. Cubert most likely would not have intended to describe a tall, thin set of stacked boxes as collectively *wide* (or *fat* or *thick*) when the distributive interpretation stands as a ready alternative. However, the reverse holds for *narrow* and the other negative width adjectives: stacking the boxes on top of each other both increases the predictability of their collective width and focuses attention on their narrowness, that is, on the truth of the collective interpretation.

The results of the current experiment thus add nuance to our claims about contextual predictability. We continue to find that collective interpretations depend on the contextual predictability of collective properties, such that increasing predictability by regularizing arrangements increases rates of collective interpretations. However, to increase predictability one must regularize the specific dimension under discussion (e.g., the height dimension for *tall* but not *long*, or the width dimension for *narrow* but not *deep*). Moreover, we have observed that contextual predictability interacts with yet another aspect of the pragmatic calculus that influences interpretation choice, namely the probability that a specific interpretation would truthfully describe a state of affairs. When an interpretation appears unlikely to be true (e.g., describing a tall stack of boxes as collectively short), listeners are unlikely to attribute that interpretation to speakers' utterances. We next formalize the role of context in the interpretation of plural predication, using tools from probabilistic modeling.

6. Modeling plural predication

We aim to formalize the pragmatic account described so far, especially the role of contextual predictability and speaker knowledge in ambiguity resolution. We will base our analysis on the Rational Speech Act (RSA) approach (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Goodman & Frank, 2016). Within the RSA framework, language understanding is modeled as a social reasoning process: the listener interprets an utterance by reasoning about a cooperative speaker who is trying to inform a naive listener about some state of affairs.

In several recent RSA models, the pragmatic listener is extended to reason over factors needed to fully fix meaning, in addition to the situation itself. That is, some aspect of the literal interpretation model is "lifted" into the inference performed by the pragmatic reasoner. This sort of lifted variable RSA model has proven useful in accounts of gradable adjective semantics (where the lifted variable is the adjective's degree threshold; Lassiter & Goodman, 2013), specificity implicatures (where the lifted variable determines the choice of lexica; Bergen, Goodman, & Levy, 2012), and non-literal meaning (where the lifted variable concerns the dimension along which the speaker intends to communicate, or the question under discussion; Kao, Wu, Bergen, & Goodman, 2014); see also Goodman and Lassiter (2015) for an overview. Our model treats the two underlying senses (collective and distributive) as vague expressions that have an underspecified threshold semantics—following Lassiter and Goodman (2013), we lift the corresponding threshold variables to be resolved by the pragmatic listener.

From Goodman and Stuhlmüller (2013), we borrow a treatment of the speaker's knowledge. The speaker has a basic motivation to inform the listener about the true state of the world; he accounts for his uncertainty about this true state by choosing an utterance according to the *expected informativity* given his belief distribution. The pragmatic listener knows this, but does not know what private observations the speaker has had—she takes this uncertainty into account when evaluating an utterance.

There are two significant innovations in the current model. First, we treat ambiguity resolution via a lifted variable: the pragmatic reasoner infers which sense (collective or distributive) of the plural predication was intended by the speaker. Second, we explore the effects of contextual noise that differentially affects the two interpretations. That is, we allow that there are uncertain contextual factors that affect evaluation of the collective property but not the distributive. In the formal exposition that follows, we will be brief about the pre-existing model features (lifted threshold variables, speaker knowledgeability) and more expansive about the new aspects.

6.1. The model

We take states of the world, $s \in S$, to consist of a collection of entities, each of which has some individual degree: for each $x \in s$, $d(x) \in \mathbb{R}$ is the relevant property. We assume a simple truth-functional literal semantics, wherein a plural predication utterance u (e.g., *the boxes are big*) denotes a mapping from states of the world to truth values.[21] We parameterize the truth function, $[\![u]\!]^{v,c,\theta_c,\theta_d} : S \rightarrow$ Bool, so that it depends on context, c, on utterance interpretation, v, and on two gradable thresholds, θ_c and θ_d. We will consider three alternative utterances: an unambiguously distributive (e.g., *the boxes each are big*), an unambiguously collective (e.g., *the boxes together are big*), and an ambiguous utterance (e.g., *the boxes are big*).

(8) *Literal semantics*:

a. $[\![\texttt{distrib}]\!]^{\theta_d} = \lambda s.\, \forall x \in s\, [d(x) > \theta_d]$

b. $[\![\texttt{coll}]\!]^{c,\theta_c} = \lambda s.\, [c + \sum_{x \in s} d(x) > \theta_c]$

c. $[\![\texttt{amb}]\!]^{v,c,\theta_d,\theta_c} =$ if v $[\![\texttt{distrib}]\!]^{\theta_d}$, else $[\![\texttt{coll}]\!]^{c,\theta_c}$

The unambiguous distributive utterance, (8a), is a universal quantification over a vague scalar meaning: each object's degree must exceed the distributive threshold θ_d. The unambiguous collective utterance, (8b), is again a vague scalar meaning: the collective degree must exceed the collective threshold θ_c. However, the collective degree captures two assumptions. First, we follow the work on collective semantics in assuming that the appropriate aggregation is based on a sum (e.g., Scha, 1984): the total degree of the collection. Second, we assume that the computation of this total degree may depend on contextual factors c (such as the particular arrangement of the objects); we treat c as a simple additive variable that can distort the total degree of the state—an approximation of the many ways that context could affect the collective property (e.g., the variability in collective height introduced by mentally stacking boxes in different arrangements). Finally, the ambiguous utterance, (8c), is governed by the variable v: depending on v it is either collective or distributive.[22]

The literal listener L_0 has prior uncertainty about the context, $P(c)$, and the true state, $P(s)$, and otherwise updates beliefs about s by conditioning on the meaning of u:

[21] We assume truth values are coerced to numbers where appropriate: *true* is 1 and *false* is 0.

[22] We ignore so-called "cumulative" construals of plural predication, which arise for sentences with sequences of noun phrases (e.g., *600 Dutch firms use 5,000 American computers*; Scha, 1984).

$$P_{L_0}(s|u, v, \theta_d, \theta_c) \propto [\![u]\!]^{v,c,\theta_d,\theta_c}(s) \cdot P(c) \cdot P(s)$$

The distribution on collective-context, $P(c)$, will determine how predictable the collective property is compared to the distributive: higher entropy $P(c)$ will correspond to noisier collective interpretations. The other interpretation variables (v, θ_d, θ_c) are lifted so that they will be actively reasoned about by the pragmatic listener. Put simply, the pragmatic listener resolves the interpretation of an ambiguous utterance and fixes the appropriate thresholds for vague gradable predicates.

The speaker, S_1, chooses an utterance u that would most effectively communicate some state s to a literal listener L_0, increasing utterance informativity by decreasing surprisal. In other words, the speaker's utility function U_{S_1} minimizes the effort L_0 would need to arrive at s from u, while being efficient (i.e., minimizing utterance cost, $C(u)$):

$$U_{S_1}(u; s, v, \theta_d, \theta_c) = \log(L_0(s|u, v, \theta_d, \theta_c)) - C(u)$$

To capture the speaker's epistemic state, we follow Goodman and Stuhlmüller (2013) by assuming the speaker has a private belief distribution about the state of the world, $P(s|o)$, that depends on some observations o of the true world. The speaker then selects an utterance u to convey information about the likely state s that generated the observation o:

$$P_{S_1}(u|o, v, \theta_d, \theta_c) \propto \exp(\alpha \mathbb{E}_{P_a(s|o)}[U_{S_1}(u; s, v, \theta_d, \theta_c)])$$

We will assume, based on our experimental setup, that the speaker has one of two modes of *access* to the world state: if $a = \texttt{full}$, then the speaker has observed the full state, in which case $P_{\texttt{full}}(s|o) = \delta_{s=o}$ is a delta distribution on the true (observed) state; if $a = \texttt{sum}$, then the speaker has only observed the sum (e.g., observing the total weight of some boxes by lifting them together), in which case the speaker considers all of the possible states that could have led to the sum observation, $P_{\texttt{sum}}(s|o) \propto P(s)\delta_{\sum_{x \in s} d(x)=o}$.

At the top level of inference, the pragmatic listener L_1 interprets the speaker's utterance u to jointly infer the state s and the free interpretation variables: the ambiguous utterance's intended interpretation v, and the thresholds θ_d, θ_c. By Bayes' rule:

$$P_{L_1}(s, v, \theta_d, \theta_c|u, a) \propto P_{S_1}(u|o(a, s), v, \theta_d, \theta_c) \cdot P(s) \cdot P(v) \cdot P(\theta_d, \theta_c)$$

Here $o(a, s)$ is a deterministic function that reads off the observation the speaker would have had given an access mode and a hypothesized true state.

The plural predication is both ambiguous (v) and vague (θ_c, θ_d); the listener reasons about likely interpretations given an intuitive understanding of speakers and prior knowledge of the world (i.e., which states and interpretations are *a priori* likely). This inference is guided by reasoning about how informative each utterance would have been with different interpretations. Informativity is in turn influenced by semantic factors including the variability of context, c, that may differentially affect the different interpretations.

6.2. Model results

To generate predictions, we need to fix various parameter settings within the model; we aim to make the simplest choices possible, focussing on qualitative predictions.[23] Each state, s, consists of three objects, each of which has degree 2, 3, or 4—these degrees correspond to object size, weight, or height. The prior on states $P(s)$ was taken to be uniform. The distributive threshold θ_d was uniformly 2, 3, or 4 (the possible object degrees). The collective threshold θ_c was uniformly drawn from the possible individual and total degrees, $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$.[24] We take $P(v)$ to be Bernoulli with a single free parameter (i.e., the prior bias toward collective interpretations); this parameter was set to 0.8. We take the cost $C(u)$ to be 2 for the ambiguous utterance and 3 for both of the unambiguous utterances—a reasonable assumption from the point of view of production, given that unambiguous utterances contain an extra word (i.e., the "surface-level" interpretation cues *each* or *together*; Syrett & Musolino, 2016). We also include a null utterance with cost 1; the null utterance corresponds to saying nothing at all and signaling the listener to rely wholly on prior beliefs. The parameter $\alpha > 0$ controls the speaker's optimality; we set α to 10. Importantly, the qualitative patterns reported below are robust to changes in these parameters and assumptions.

Our central claim surrounds the effect of $P(c)$, which controls the predictability of the collective interpretation. To evaluate its role, we generated model predictions for distributions of varying entropy; c was drawn from a normal distribution centered around 0. The variance of this distribution controls the amount of collective interpretation noise. We begin by comparing the effects of `no` ($\sigma = 0.01$), `low` ($\sigma = 1$), `mid` ($\sigma = 2$), and `high` ($\sigma = 3$) collective noise. In Fig. 11, we plot the probability that the interpretation-resolving variable v is `collective` after hearing the `ambiguous` utterance, marginalizing the other variables from the pragmatic listener L_1. In other words, we plot the probability of a collective interpretation depending on collective noise and speaker access.

The qualitative predictions of our plural predication model match the hypotheses and experimental results that motivated the model. Most striking is the monotonic decrease in probability of collective interpretations as collective noise increases. We now have a formal understanding as to why this effect holds: the noisier the collective interpretation, the less likely the listener L_0 is to correctly resolve the question under discussion, namely the properties of the named boxes s. Given that the speaker S_1's goal is to successfully communicate s to L_0, the noisier an interpretation, the less useful it is at achieving this goal; the speaker is therefore less likely to use the ambiguous utterance, and as a result the pragmatic listener L_1 is less likely to infer that the ambiguous utterance, when used, meant the noisy collective meaning.

The model also successfully predicts the qualitative increase in the probability of a collective interpretation when the speaker's access to individual object properties is limited (red vs. blue bars in Fig. 11).[25] L_1 tracks S_1's epistemic state, and therefore knows whether the speaker has full or partial knowledge. In the case of partial (i.e., sum) knowledge, a distributive interpretation lacks epistemic support and therefore becomes less likely; as a result, collective interpretations become more likely.

To compare the quantitative predictions of our model with the results of Expt. 3, we would need to fit the variance of $P(c)$ for each predicate and each arrangement condition (i.e., each of the red bars in Fig. 9). But with only 12 data points by which to fit these 12 parameters, we would not want to claim that the model uniquely, quantitatively predicts the data. Future work will need to experimentally measure the appropriate noise values and ideally do parametric manipulations of the conditions.

7. General discussion

We began with an old observation about collective interpretations for gradable predicates with plural subjects: some predicates

[23] The full model is viewable online at http://forestdb.org/models/plural-predication.html.

[24] Model predictions remain qualitatively the same if θ_c is drawn instead from only the possible total degrees, $\{6, 7, 8, 9, 10, 11, 12\}$.

[25] For interpretation of color in 'Fig. 11', the reader is referred to the web version of this article.

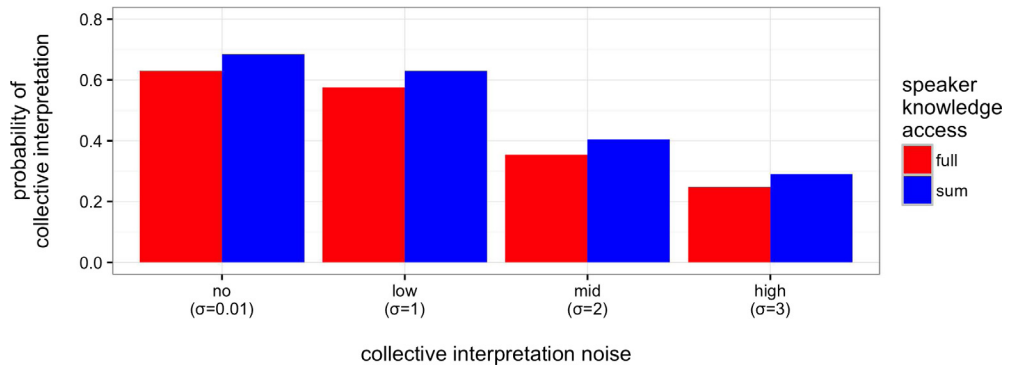

Fig. 11. Model predictions for the probability of the `ambiguous` utterance receiving a collective interpretation as a function of collective interpretation noise and speaker knowledge access.

admit collective interpretations, while others strongly (or totally) resist them. The puzzle is made even more pronounced by the lack of conceptual barriers to the collective construals of these predicates (cf. the truly distributive predicates like *be a man* or *have blue eyes*). This phenomenon, dubbed "stubborn distributivity" after those gradable predicates that ostensibly refuse collective interpretations, has been documented and described, even characterized in terms of semantic constraints. Still, an explanation of the phenomenon has proven elusive. Here, we offer an explanation: stubbornly distributive predicates are so because the collective properties they name are unpredictable, or unstable, in most contexts; this unpredictability results in a noisy collective interpretation, something speakers and listeners recognize as ineffective for communicating efficiently about their world. On this pragmatic account, stubborn distributivity does not form a distinct category; rather, there is a gradient of stubbornness. Moreover, the extent that collective interpretations are drawn should be sensitive to contextual factors.

We found evidence for this pragmatic explanation over the course of four experiments. Perhaps most problematic for a categorical account of stubborn distributivity were the results of Expt. 2, where we showed that predicates do not form neat groupings according to whether they are stubbornly distributive; stubbornness indeed is gradient. It further depends on many aspects of the context of predication. Here we focused on two such aspects: the contextual predictability of collective properties and speakers' access to knowledge.

After manipulating the regularity of the predication context directly in Expt. 3, we found that more regular contexts yielded greater rates of collective paraphrase endorsement. We thus confirmed our hypothesis: increasing contextual predictability of collective properties—by increasing the regularity of the contexts themselves—increases rates of collective interpretations. In other words, rates of collective interpretation were shown to be malleable, in this case influenced by the predictability of the property named in the specific context of predication.

We similarly confirmed our hypothesis regarding speaker knowledge: in Expts. 1 and 3, we manipulated the discourse context in such a way as to plausibly limit the speaker's access to knowledge of individual properties. We found that when speakers lacked epistemic support for a distributive interpretation, collective interpretations became more likely; speakers are less likely to intend an interpretation for which they lack epistemic support.

In Section 6, we formalized our hypotheses with a computational model of pragmatic dissambiguation that gives a precise role for contextual noise and speaker knowledge. This pragmatic model indeed predicts that the acceptability of collective readings should depend on predictability and speaker knowledge. Formalizing the pragmatic approach is an important step in validating it as a coherent theory of plural predication and stubborn distributivity.

7.1. Implications for semantic theories of plural predication

Recall that current theories of stubborn distributivity describe it as a constraint on the sort of subjects certain predicates may compose with. Given the prohibition on plural subjects in their basic denotations, stubbornly distributive predicates simply cannot deliver collective interpretations. At the very least, the results of our experiments demonstrate that stubborn distributivity should not (indeed, cannot) manifest in terms of an all-out prohibition against collective interpretations. Even unabashedly stubbornly distributive predicates like *big* yield greater rates of collective interpretations when context supports them; we claim it is the way that context supports collective interpretations that provides the key to understanding where stubborn distributivity comes from in the first place: contextual predictability.

Complaisantly collective predicates like *heavy* or *expensive* do not require support from context to ensure that their collective properties are predictable. The properties of collective weight and price are inherently stable in context; there exists but one way to evaluate these properties, and they persist despite changing physical arrangement, speaker perspective, etc. Not so for the collective properties named by stubbornly distributive predicates like *big* or *tall*, which name properties of physical extent. Owing to uncertainty in evaluation strategy, compounded by definitional dependence on physical arrangement, these collective properties are relatively unpredictable, unstable, and as a result difficult for speakers and listeners to coordinate on in the absence of additional contextual or linguistic cues. Grammar might encode stubborn distributivity as a preference against plural arguments for certain predicates (or as a preference for plural arguments with fixed physical arrangements), but here we suggest a simpler explanation for this preference: stubbornly distributive predicates name collective properties that are difficult to accurately communicate. Increasing the contextual predictability of these properties decreases this difficulty and thereby increases the utility and likelihood of collective interpretations.

Further evidence for the non-trivial role of contextual predictability in collective predication comes from data demonstrating that one and the same predicate receives varying rates of collective interpretations depending on the noun that serves as its subject. Looking at naturally-occurring examples of plural predication, we identified the most frequent subject nouns for the stubbornly distributive predicates *big* and *small* in a plural predication sentence frame. These predicates generally resisted collective interpretations, as we would expect if they were stubbornly

distributive, but certain subject nouns yielded greater rates of collective interpretations. Crucially, these seemed to be subject nouns that suggested relatively specific collective properties. Again, contextual predictability of the collective property determines availability of collective interpretations.

Note that subject nouns already played a central role in the disambiguation of plural predications. Schwarzschild (2011) observed that physically aggregating a plurality using an atomizing noun like *pile* or *stack* greatly increases the availability of collective interpretations. He observed a similar, though opposite effect for mass nouns, which lack any information about physical arrangement. Again, semantics might attend to the atomic status of the resulting noun phrase in its evaluation of stubborn distributivity (e.g., *the pile of boxes* names a single atom, while *the boxes* most likely does not), but pragmatics provides us with an explanation for why these collectivizing nouns seem to unlock the collective interpretation. Just as our contextual manipulation increased contextual predictability of collective size or height by increasing the likelihood of encountering a plurality of boxes in a specific physical arrangement, so too do atomizing nouns. The arrangement of a *stack* or *pile* of boxes is much more regular across contexts than an abstract collection thereof. It is this regularity introduced by the atomizing noun that permits an otherwise elusive collective interpretation; the pragmatic mechanism of determining the reading remains the same.

To understand this mechanism, namely how contextual predictability affects the calculus of interpretation choice, we modeled its contribution explicitly within a computational model of plural predication. Both the formal analysis and experimental results operate at the level of whole utterances. They leave open several questions of compositional semantics. Most notably, we assume that a plural predication is ambiguous between collective and distributive readings. The locus of this ambiguity could be in the predicate, in the noun phrase, or only in the composition of the two. For instance, it may be that the noun phrase can be construed as either a plurality (e.g., some boxes) or an aggregate (e.g., a single collection), giving rise to the ambiguity in plural predication. Future research will be needed to pin down the compositional source of the ambiguity; having a theory of how the ambiguity is resolved in context is likely to help.

7.2. Implications for formal models of pragmatics

Our explanation for stubborn distributivity relies on a sophisticated process of recursive reasoning involving listeners and speakers. Thanks to recent advances in computational cognitive science (in our case, simulation-based probabilistic programs; Goodman, Tenenbaum, & Gerstenberg, 2015), we are able to move beyond simply describing this reasoning process by implementing a formal pragmatic model that makes quantitative predictions. By considering utterance semantics within an articulated model of pragmatics, what began as a description of the phenomenon—certain predicates resist collective interpretations—now has an explanation: speakers avoid interpretations that are unlikely to correctly resolve the question under discussion (e.g., collective interpretations of stubbornly distributive predicates), and listeners know this about speakers, so they avoid inferring interpretations that speakers avoid using.

Our model was formulated within the Rational Speech Act framework, and explores two extensions that may have broader consequences. First, we used the "lifted variable" approach to formulate ambiguity resolution as an active pragmatic process. Ambiguity is endemic to natural language (for a recent discussion, see Piantadosi, Tily, & Gibson, 2012); taken out of context, nearly everything we say lacks its intended meaning. Applying the pragmatic approach to other cases of ambiguity may be a fruitful way to understand complex interactions between semantics, context, and world-knowledge in cases of multiple potential meanings.

Second, we highlighted the role that context-specific predictability, or noise, in computing meaning can have on pragmatic interpretation. Interlocutors are finely tuned to the pragmatic usefulness of expressions in language; unavoidable noise in interpretation can make expressions less useful, and hence dis-preferred. We have applied this mechanism to disambiguition in plural predication, but it is likely to have an impact throughout language.

8. Conclusion

We have argued that understanding plural predication, and notably the phenomenon of stubborn distributivity, arises from an interaction between the rational reasoning processes of language users and qualities of the collective properties they name. Speakers and listeners require more coordination to align their beliefs about properties that are less predictable across contexts (e.g., collective size or shape, which might change as arrangements change). As a result, they require more support from context to accurately discuss these potentially variable, uncertain properties. This support might come in the form of words that increase contextual predictability (e.g., *stack*, which regularizes otherwise variable physical arrangements) or from increasing the regularity of the context itself (e.g., by ensuring that sets realize a common physical arrangement). As collective properties become more predictable, collective interpretations become more informative, more useful, and therefore more likely.

Appendix A. Full mixed effects logistic regression models from Expt. 1

Table A.1 presents model coefficients for the full mixed logistic regression model used in the analysis of disambiguating paraphrases in Expt. 1. The model predicts collective referent choice from fixed effects for disambiguating utterance and its interaction with predicate, together with trial order. The model includes random by-participant and by-scenario intercepts. The fixed effects predictors were centered before analysis.

Table A.2 presents model coefficients for the full mixed logistic regression model used in the analysis of bare utterances in Expt. 1. The model predicts collective referent choice from fixed effects for predicate, scenario, and trial, as well as random by-participant intercepts and slopes grouped by trial. Fixed effects predictors were centered before analysis.

Table A.1
Full logistic regression model from paraphrase analysis of Expt. 1.

	Coef β	SE(β)	z	p
Intercept	−0.25	0.25	−0.97	0.33
Utterance	4.99	0.82	6.08	<0.01
Trial	0.07	0.07	1.00	0.32
Utterance:Pred-heavy	−1.39	0.91	−1.52	0.13
Utterance:Pred-tall	−0.63	0.95	−0.66	0.51

Table A.2
Full logistic regression model from bare utterance analysis of Expt. 1.

	Coef β	SE(β)	z	p
Intercept	−4.65	1.49	−3.12	<0.01
Pred-heavy	3.05	1.23	2.48	<0.05
Pred-tall	3.63	1.29	2.82	<0.01
Trial	−0.17	0.20	−0.84	0.40

Appendix B. Full mixed effects linear regression models from Expt. 2

Table B.1 presents model coefficients for the full mixed linear regression models used in the analyses of the 8 predicates occurring with more than one subject noun in Expt. 2a. The models predict collective endorsement ratings from fixed effects for subject noun, as well as random by-participant intercepts. For each predicate, the subject noun with the highest mean collective endorsement rating served as the reference level.

Table B.2 presents model coefficients for the full mixed linear regression models used in the analyses of subject nouns in Expt. 2b. The models predict collective endorsement ratings for each of the three predicates (*big*, *heavy*, *tall*) from fixed effects for subject noun, as well as random by-participant intercepts. For each predicate, the subject noun with the highest mean collective endorsement rating served as the reference level.

Table B.1
Full linear regression models from subject noun analysis of Expt. 2a.

	Coef β	SE(β)	t	p
Bright				
Intercept-eyes	0.75	0.07	11.46	<0.01
Noun-colors	−0.19	0.08	−2.26	<0.05
Closed				
Intercept-doors	0.60	0.07	8.61	<0.01
Noun-shops	−0.16	0.09	−1.87	0.07
Friendly				
Intercept-natives	0.73	0.05	13.41	<0.01
Noun-people	−0.05	0.06	−0.93	0.37
Full				
Intercept-streets	0.68	0.06	11.53	<0.01
Noun-papers	−0.11	0.10	−1.19	0.24
Noun-pubs	−0.23	0.08	−2.91	<0.01
Guilty				
Intercept-defendants	0.68	0.05	12.51	<0.01
Noun-men	−0.02	0.08	−0.21	0.84
Open				
Intercept-eyes	0.65	0.07	9.73	<0.01
Noun-gates	0.01	0.09	0.08	0.94
Noun-doors	−0.03	0.08	−0.40	0.69
Noun-curtains	0.00	0.09	−0.03	0.98
Noun-pubs	−0.11	0.09	−1.29	0.20
Noun-windows	−0.15	0.08	−1.77	0.08
Quiet				
Intercept-dogs	0.71	0.07	10.72	<0.01
Noun-streets	−0.06	0.09	−0.65	0.52
Small				
Intercept-numbers	0.57	0.06	9.65	<0.01
Noun-children	−0.16	0.07	−2.34	<0.05
Noun-rooms	−0.20	0.07	−2.78	<0.01
Noun-classes	−0.28	0.07	−3.88	<0.01

Table B.2
Full linear regression models from subject noun analysis of Expt. 2b.

	Coef β	SE(β)	t	p
Big				
Intercept-waves	0.42	0.06	6.73	<0.01
Noun-rooms	−0.18	0.06	−2.96	<0.01
Noun-boys	−0.22	0.06	−3.53	<0.01
Noun-children	−0.21	0.06	−3.54	<0.01
Noun-houses	−0.23	0.06	−3.68	<0.01
Heavy				
Intercept-bags	0.63	0.07	9.63	<0.01
Noun-lids	−0.08	0.08	−1.04	0.30
Noun-trees	−0.15	0.08	−1.93	0.06
Noun-loads	−0.24	0.08	−3.15	<0.01
Noun-men	−0.25	0.08	−3.36	<0.01
Tall				
Intercept-trees	0.30	0.06	4.65	<0.01
Noun-plants	0.00	0.06	−0.07	0.95
Noun-offspring	−0.03	0.06	−0.55	0.59
Noun-buildings	−0.06	0.06	−1.00	0.32
Noun-windows	−0.09	0.06	−1.52	0.13

Appendix C. Full linear regression models from Expt. 3

Table C.1 presents model coefficients for the full linear regression models used in the analysis of Expt. 3. For each predicate, the models predict collective endorsement ratings from fixed effects for context, scenario, trial, and the interaction between context and scenario. The fixed effects predictors were centered before analysis.

Table C.1
Full linear regression models from analysis of Expt. 3.

	Coef β	SE(β)	t	p
Big				
Intercept	0.57	0.04	15.69	<0.01
Context	0.18	0.07	2.52	<0.05
Scenario	0.00	0.07	0.03	0.97
Trial	−0.03	0.04	−0.67	0.51
Context:Scenario	0.21	0.15	1.43	0.16
Heavy				
Intercept	0.61	0.04	16.47	<0.01
Context	0.11	0.07	1.43	0.16
Scenario	0.19	0.08	2.54	<0.05
Trial	0.02	0.05	0.42	0.67
Context:Scenario	−0.12	0.15	−0.83	0.41
Tall				
Intercept	0.66	0.03	19.99	<0.01
Context	0.30	0.07	4.58	<0.01
Scenario	0.07	0.07	1.11	0.27
Trial	−0.01	0.04	−0.25	0.80
Context:Scenario	0.02	0.13	0.17	0.86

Appendix D. Full linear regression models from Expt. 4

Tables D.1 and D.2 present model coefficients for the full linear regression models used in the analysis of Expt. 4. For each dimension, the models predict collective endorsement ratings from fixed effects for context and trial; for those dimensions with both positive- and negative-polarity adjectives, the models also included fixed effects of polarity and the interaction between polarity and context. The fixed effects predictors were centered before analysis.

Table D.1
Full linear regression models from the positive-polarity dimension analyses of Expt. 4.

	Coef β	SE(β)	t	p
Capacity				
Intercept	0.38	0.08	4.61	<0.01
Context	0.16	0.08	1.95	<0.06
Trial	0.00	0.01	0.75	0.46
Depth				
Intercept	0.58	0.08	7.27	<0.01
Context	0.07	0.09	0.85	0.40
Trial	−0.01	0.01	−1.71	0.09
Length				
Intercept	0.47	0.07	7.07	<0.01
Context	0.01	0.06	0.15	0.89
Trial	0.00	0.00	0.28	0.78
Weight				
Intercept	0.78	0.08	10.05	<0.01
Context	−0.04	0.07	−0.51	0.62
Trial	−0.01	0.01	−1.86	<0.07

Table D.2
Full linear regression models from the multi-polarity dimension analyses of Expt. 4.

	Coef β	SE(β)	t	p
Height				
Intercept	0.44	0.04	11.53	<0.01
Context	0.04	0.04	1.12	0.26
Polarity	0.28	0.04	7.82	<0.01
Trial	0.00	0.00	0.30	0.76
Context:Polarity	0.13	0.07	1.82	<0.07
Size				
Intercept	0.42	0.03	15.52	<0.01
Context	0.03	0.03	0.94	0.35
Polarity	0.20	0.03	7.60	<0.01
Trial	−0.00	0.00	−0.16	0.87
Context:Polarity	0.13	0.05	2.39	<0.05
Width				
Intercept	0.45	0.03	12.95	<0.01
Context	0.06	0.03	1.74	0.08
Polarity	0.06	0.03	1.79	0.07
Trial	−0.00	0.00	−0.98	0.33
Context:Polarity	−0.14	0.07	−2.02	<0.05

Appendix E. Materials from Experiment 2

Tables E.1 and E.2 present the full set of attested sentences used in Expts. 2a and 2b, respectively.

Table E.1
Full set of attested sentences from Expt. 2a.

Item	Sentence	Item	Sentence
1	The windows were dark	21	The men were guilty
2	The windows were open	22	The items were unreasonable
3	The walls were bare	23	The injuries were consistent
4	The walls were white	24	The guests were due
5	The streets were full	25	The gates were open
6	The streets were quiet	26	The facts were similar
7	The shops were closed	27	The eyes were bright
8	The shares were unchanged	28	The eyes were open
9	The rooms were small	29	The doors were closed
10	The roads were busy	30	The doors were open
11	The results were encouraging	31	The dogs were quiet
12	The results were disappointing	32	The defendants were liable
13	The results were gratifying	33	The defendants were guilty
14	The pubs were full	34	The curtains were open
15	The pubs were open	35	The conditions were appalling
16	The people were friendly	36	The conditions were difficult
17	The papers were full	37	The colors were bright
18	The omens were good	38	The classes were small
19	The numbers were small	39	The children were small
20	The natives were friendly	40	The children were young

Table E.2
Full set of attested sentences from Expt. 2b.

Item	Sentence
1	The houses were big
2	The waves were big
3	The rooms were big
4	The boys were big
5	The children were big
6	The trees were heavy
7	The men were heavy
8	The loads were heavy
9	The lids were heavy
10	The bags were heavy
11	The trees were tall
12	The offspring were tall
13	The buildings were tall
14	The windows were tall
15	The plants were tall

Appendix F. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2017.07.002.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. (2014). lme4: Linear mixes-effects models using Eigen and S4. *Journal of Statistical Software*. http://arxiv.org/abs/1406.5823. ArXiv e-print.

Bergen, L., Goodman, N. D., & Levy, R. (2012). That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society*. Wheat Ridge, CO: Cognitive Science Society.

Davies, M. (2011). *Google Books Corpus*. (Based on Google Books n-grams).

DiCiccio, T. J., & Efron, B. (1996). Boostrap confidence intervals (with discussion). *Statistical Science, 11*, 189–228.

Dixon, R. (1982). *Where have all the adjectives gone? And other essays in semantics and syntax*. Berlin: Mouton.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336*, pp. 998-998.

Gilbert, M. P. (2013). *Joint commitment: How we make the social world*. New York: Oxford University Press.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences, 20*(11), 818–829.

Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. In S. Lappin & C. Fox (Eds.), *The handbook of contemporary semantics* (2nd ed.. Wiley-Blackwell.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science, 5*, 173–184.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Laurence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 623–654). Cambridge, MA: MIT Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts*. New York: Academic Press, pp. 41–58li.

Higginbotham, J., & Schein, B. (1989). Plurals. In J. Carter & R.-M. Déchaine (Eds.). *Proceedings of NELS* (Vol. 19, pp. 161–175). Amherst, MA: Graduate Linguistics Students Association, University of Massachusetts.

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition, 57*, 129–191.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Non-literal understanding of number words. *Proceedings of the National Academy of Sciences of the United States of America, 111*, 12002–12007.

Kennedy, C. (1999). *Projecting the adjective: The syntax and semantics of gradability and comparison*. New York: Garland.

Landman, F. (1989a). Groups I. *Linguistics and Philosophy, 12*, 559–605.

Landman, F. (1989b). Groups II. *Linguistics and Philosophy, 12*, 723–744.

Landman, F. (1996). Plurality. In S. Lappin (Ed.), *Handbook of contemporary semantics* (pp. 425–457). Oxford: Blackwell.

Lasersohn, P. (1988). *A semantics for groups and events* Ph. D. thesis. The Ohio State University.

Lasersohn, P. (1990). Group action and spatio-temporal proximity. *Linguistics and Philosophy, 13*, 179–206.

Lasersohn, P. (1998). Generalized distributivity operators. *Linguistics and Philosophy, 21*, 83–93.

Lasersohn, P. N. (1995). *Plurality, conjunction and events*. Dordrecht: Kluwer Academic Publishers.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In T. Snider (Ed.). *Proceedings of SALT* (Vol. 23, pp. 587–610). CLC Publications.

Link, G. (1983). The logical analysis of plurals and mass terms. In R. Bäuerle, C. Schwarze, & A. von Stechow (Eds.), *Meaning, use, and interpretation of language* (pp. 302–323). Berlin: de Gruyter.

Link, G. (1987). Generalized quantifiers and plurals. In P. Gärdenfors (Ed.), *Generalized quantifiers* (pp. 151–180). Dordrecht: D. Reidel.

Link, G. (1998). Ten years of research on plurals – where do we stand? In F. Hamm & E. Hinrichs (Eds.), *Plurality and quantification. Studies in linguistics and philosophy* (Vol. 69, pp. 19–54). Netherlands: Springer.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative efficiency of ambiguity in language. *Cognition, 122*, 280–291.

Quine, W. V. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Scha, R. (1984). Distributive, collective and cumulative quantification. In *Truth, interpretation, and information* (pp. 131–158). Dordrecht: Foris.

Schein, B. (2017). 'And': Conjunction reduction redux. Cambridge, MA: MIT Press.

Schwarzschild, R. (1994). Plurals, presuppositions and the sources of distributivity. *Natural Language Semantics, 2*(3), 201–248.

Schwarzschild, R. (1996). *Pluralities*. Dordrecht: Kluwer Academic Publishers.

Schwarzschild, R. (2011). Stubborn distributivity, multiparticipant nouns and the count/mass distinction. In S. Lima, K. Mullin, & B. Smith (Eds.). *Proceedings of NELS* (Vol. 39, pp. 661–678). Amherst, MA: Graduate Linguistics Students Association, University of Massachusetts.

Syrett, K. (2015). Mapping properties to individuals in language acquisition. In *BUCLD 39 proceedings*. Cascadilla Press.

Syrett, K., & Musolino, J. (2013). Collectivity, distributivity, and the interpretation of plural numerical expressions in child and adult language. *Language Acquisition, 20*(4), 259–291.

Syrett, K., & Musolino, J. (2016). All together now: Disentangling semantics and pragmatics with *together* in child and adult language. *Language Acquisition, 23* (2), 175–197.

Vázquez Rojas Maldonado, V. (2012). *The syntax and semantics of Purépecha noun phrases and the mass/count distinction* Ph. D. thesis. New York University.

Zhang, N. N. (2013). *Classifier structures in Mandarin Chinese*. Berlin: Mouton de Gruyter.