Remarkable features: Using descriptive contrast to convey and infer typicality

Claire Augusta Bergey[1]

[1] The University of Chicago

Remarkable features: Using descriptive contrast to convey and infer typicality

An utterance can say much more about the world than its literal interpretation might suggest. For instance, if you hear a colleague say "We should hire a female professor," you might infer something about the speaker's goals, the makeup of a department, or even the biases of a field—none of which is literally stated. These inferences depend on recognition that a speaker's intended meaning can differ from the literal meaning of their utterance, and the process of deriving this intended meaning is called *pragmatics*. Frameworks for understanding pragmatic inference posit that speakers tend to follow general principles of conversation—for instance, that they tend to be relevant, brief, and otherwise helpfully informative (Clark, 1990; Grice, 1975; Sperber & Wilson, 1986). When a speaker deviates from these principles, a listener can reason about the alternative utterances the speaker might have said and infer some intended meaning that goes beyond the literal meaning of their utterance.

Beyond enriching the interpretation of utterances whose literal meaning is known, pragmatic inference is a potentially powerful mechanism for learning about new words and concepts. People can learn the meanings of words by tracking associations between word use and present objects alone (Yu & Smith, 2007), but reasoning about a speaker's intended meaning—not just relating the words they say to objects in the environment—may support more rapid and more accurate learning (Frank, Goodman, & Tenenbaum, 2009). For example, Akhtar, Carpenter, and Tomasello (1996) showed that young children can infer the meaning of a new word by using the principle that people tend to remark on things that are new and interesting to them. In this study, an experimenter leaves the room and a new toy emerges in her absence; once she comes back, the toy is familiar to the child but not to the experimenter. When she uses a novel name, "gazzer," the child can infer that the word refers to the toy that is novel to the experimenter, and not to other toys the experimenter had already seen. Experiments with adults show that they too can use general principles of informativeness to infer a novel referent's name (Frank & Goodman, 2014).

One potential pragmatic tool for learning about referents is contrastive inference from description. To the extent that communicators strive to be minimal and informative, description should discriminate between the referent and some relevant contrasting set. This contrastive inference is fairly obvious from some types of description, such as some postnominal modifiers: "The door with the lock" clearly implies a contrasting door without one (Ni, 1996). The degree of contrast implied by more common descriptive forms, such as prenominal adjectives in English, is less clear: speakers do not always use prenominal adjectives minimally, often describing more than is needed to establish reference (Engelhardt, Barış Demiral, & Ferreira, 2011; Mangold & Pobel, 1988; Pechmann, 1989). Nevertheless, Sedivy, Tanenhaus, Chambers, and Carlson (1999) showed that people can use these inferences to resolve referential ambiguity in familiar contexts. When asked to "Pick up the tall cup," people directed their attention more quickly to the target when a short cup was present, and did so in the period before they heard the word "cup." Because the speaker would not have needed to specify "tall" unless it was informative, listeners were able to use the adjective to direct their attention to a tall object with a shorter counterpart. Subsequent work using similar tasks has corroborated that people can use contrastive inferences to direct their attention among familiar referents (Aparicio, Xiang, & Kennedy, 2016; Ryskin, Kurumada, & Brown-Schmidt, 2019; Sedivy, 2003).

But what if you didn't know the meaning of the key words in someone's utterance—could you use the same kind of contrastive inferences to learn about new words and categories? Suppose a friend asks you to "Pass the tall dax." Intuitively, your friend must have said the word "tall" for a reason. One possibility is that your friend wants to distinguish the dax they want from another dax they do not. In this case, you might look around the room for two similar things that vary in height, and hand the taller one to them. If, alternatively, you only see one object around whose name you don't know, you might draw a different inference: this dax might be a particularly tall dax. In this case, you might think your friend used the word "tall" for a different reason–not to distinguish the dax they want from other daxes around you, but to distinguish the dax they want from other daxes in the world. This would be consistent with data from production studies, in which people tend to describe atypical features more than they describe typical ones (Mitchell, Reiter, & Deemter, 2013; Rubio-Fernández, 2016a; Westerbeek et al., 2015a). For instance, people almost always say "blue banana" to refer to a blue banana, but almost never say "yellow banana" to refer to a yellow one. In each of these cases—when distinguishing the dax from other referents nearby, or from daxes in general—you would have used a pragmatic inference to learn something new about the category of daxes.

This dissertation will explore the ways in which people can learn about new words and categories from contrastive inference, with an eye toward understanding how contrastive inference could help children learn about language and the world it describes. To set the stage for understanding how listeners use contrastive inference, we first need to establish that speakers use adjectives in informative ways.

In Chapter 1, we investigate whether people tend to use adjectives to remark on the atypical features (e.g., "the purple carrot") rather than the typical features (e.g., "the [orange] carrot") of things. In a corpus study of caregivers' speech, we show that caregivers tend to mention atypical rather than typical features of things when speaking to their children. We also preliminarily ask whether adults speaking to other adults tend to remark on atypical features rather than typical ones in a large naturalistic corpus, extending findings from reference game tasks in the lab (Mitchell et al., 2013; Rubio-Fernández, 2016a; Westerbeek et al., 2015a), and propose to investigate this question in more depth. In an analysis using language models, we also examine whether it is possible to learn about the typical features of things without pragmatic inference, using the statistical patterns within language alone. To do this, we examine whether two language models that use word co-occurrence to represent word meaning, word2vec and BERT, represent nouns as more similar to their typical adjectives than their atypical adjectives. We find that they do not: likely because they use associative methods to represent word meaning while their input tends to highlight atypical features, these models represent the relationship between nouns and adjectives poorly. We discuss implications for children's word learning as well as for language modeling.

In Chapter 2, we will establish that adults can use contrastive inferences to learn about a new category's feature distribution. People use adjectives for multiple communicative purposes: in some cases, an adjective is needed to pick out one object among others in the immediate environment (e.g., "the tall cup" contrasts with a nearby shorter cup, but is not especially tall); in others, it marks atypicality (e.g., "the tall cup" is taller than most cups in general). In this chapter, we use two experiments with adults to show that people can use contrastive inferences to learn about a new category's feature distribution. People observe instances of novel categories and hear them described (e.g., "Pass me the [green] toma"), and then judge the prevalence of the relevant feature (e.g., how common it is for tomas to be green). People infer that mentioned features are less prevalent than unmentioned ones, and do so even when the feature had to be mentioned to establish reference. We use a model in the Rational Speech Act (RSA) framework to capture people's judgments, finding that their judgments reflect graded consideration of both reference and conveying typicality as purposes of using an adjective.

In Chapter 3, we propose to test whether children are able to use contrastive inferences to learn about the feature distributions of new categories. To do this, we will examine children's contrastive inferences about two domains: novel objects and novel social groups. We will test whether children make the inference that, for example, mentioning that a certain group member is smart, kind or strong implies that other group members are less likely to have those traits. We will also test whether mentioning that an object is round, tall or furry prompts children to make a similar inference. Investigating both of these domains allows us to understand how children come to make adult-like contrastive inferences and whether these inferences differ across categories about which children might receive more or less direct evidence. We discuss the implications of this kind of inference for children's learning given the descriptions they hear from caregivers, and the potential unintended consequences of remarking on individuals' traits for children's learning about social groups.

Chapter 1: People talk more about atypical than typical features of things

Children learn a tremendous amount about the structure of the world around them in just a few short years, from the rules that govern the movement of physical objects to the hierarchical structure of natural categories and even the relational structures among social and cultural groups (Baillargeon, 1994; Legare & Harris, 2016; Rogers & McClelland, 2004). Where does the information driving this rapid acquisition come from? Undoubtedly, a sizeable portion comes from direct experience observing and interacting with the world (Sloutsky & Fisher, 2004; Stahl & Feigenson, 2015). But another important source of information comes from the language people use to talk about the world (Landauer & Dumais, 1997; Rhodes, Leslie, & Tworek, 2012). How similar is the information from children's direct experience to the information available in the language children hear?

Two lines of work suggest that they may be surprisingly similar. One compelling area of work is the comparison of semantic structures learned by congenitally blind children to those of their sighted peers. In several domains that would seem at first blush to rely heavily on visual information, such as verbs of visual perception (e.g., *look*, *see*), blind children and adults make semantic similarity judgments that mirror their sighted peers (Bedny, Koster-Hale, Elli, Yazzolino, & Saxe, 2019; Landau, Gleitman, & Landau, 2009). A second line of evidence supporting the similarity of information in perception and language is the broad success of statistical models trained on language alone in approximating human judgments across a variety of domains (Landauer & Dumais, 1997; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Even more compellingly, models trained on both language usage and perceptual features for some words can infer the perceptual features of linguistically related words entirely from the covariation of language and perception (Johns & Jones, 2012).

Still, there is reason to believe that some semantic features may be harder to learn from language than these data suggest. This is because we rarely use language merely to provide running commentary on the world around us; instead, we use language to talk about things that diverge from our expectations or those of our conversational partner (Grice, 1975). People tend to avoid being over- or under-informative when they speak. In particular, when referring to objects, people are informative with respect to both the referential context and the typical features of the referent (Rubio-Fernández, 2016b; Westerbeek et al., 2015b). People tend to refer to an object that is typical of its category with a bare noun (e.g., calling an orange carrot "a carrot"), but often specify when an object has an atypical feature (e.g, "a purple carrot"). Given these communicative pressures, naturalistic language statistics may provide surprisingly little evidence about what is typical (Willits, Sussman, & Amato, 2008).

If parents speak to children in this minimally informative way, children may be faced with input that emphasizes atypicality in relation to world knowledge they do not yet have. For things like carrots—which children learn about both from perception and from language—this issue may be resolved by integrating both sources of information. Likely almost all of the carrots children see are orange, and hearing an atypical exemplar noted as a "purple carrot" may make little difference in their inferences about the category of carrots more broadly. But for things to which they lack perceptual access—such as rare objects, unfamiliar social groups, or inaccessible features like the roundness of the Earth—much of what they learn must come from language (Harris & Koenig, 2006). If language predominantly notes atypical features rather than typical ones, children may overrepresent atypical features as they learn the way things in the world tend to be.

On the other hand, parents may speak to children differently from the way they speak to other adults. Parents' speech may reflect typical features of the world more veridically, or even emphasize typical features in order to teach children about the world. Parents alter their speech to children along a number of structural dimensions, using simpler syntax and more reduplications (Snow, 1972). Their use of description may reflect similar alignment to children's abilities by emphasizing typical feature information children are still learning.

We examine the typicality of adjectives in a large, diverse corpus of parent-child interactions recorded in children's homes to ask whether parents talking to their children tend to use adjectives predominantly to mark atypical features. We find that they do: Parents overwhelmingly choose to mention atypical rather than typical features. We also find that parents use adjectives differently over the course of children's development, noting typical features more often to younger children. We then ask whether the co-occurrence structure of language nonetheless captures typicality information by training vector space models on child-directed speech. We find that relatively little typical feature information is represented in these semantic spaces.

Adjective typicality

In order to determine whether parents use adjectives mostly to mark atypical features of categories, we analyzed caregiver speech from a large corpus of parent-child interactions. We extracted a subset of adjective-noun combinations that co-occurred, and asked a sample of Amazon Mechanical Turkers to judge how typical the property described by each adjective was for the noun it modified. We then examined both the broad features of this typicality distribution and the way it changes over development. Our theoretical hypotheses, statistical models, sample size, and exclusion criteria were pre-registered on the Open Science Framework (https://osf.io/ypdzv/).

Corpus

We used data from the Language Development Project, a large-scale, longitudinal corpus of parent-child interactions recorded in children's homes. Families were recruited to be representative of the Chicagoland area in both socio-economic and racial composition (Goldin-Meadow et al., 2014). Recordings were taken in the home every 4 months from when the child was 14 months old until they were 58 months old, resulting in 12 timepoints. Each recording was of a 90-minute session in which parents and children were free to behave and interact as they liked.

Our sample consisted of 64 typically-developing children and their caregivers with data from at least 4 timepoints (*mean* = 11.3 timepoints). Together, this resulted in a total of 641,402 distinct parent utterances.

Stimulus Selection

From these utterances, we extracted all of the nouns (using human-coded part of speech tags) resulting in a set of 8,150 total nouns. Because of our interest in change over development, we considered only nouns that appeared at least once every 3 sessions (i.e. at least once per developmental year). This yielded a set of 1,829 potential target nouns used over 198,014 distinct utterances.

We selected from the corpus all 35,761 distinct utterances containing any of these nouns and any word tagged as an adjective. We considered for analysis all adjective-noun pairs that occurred in any utterance (e.g., utterances with one noun and three adjectives were coded as three pairs) for a total of 18,050 distinct pairs. This set contained a number of high-frequency idiomatic pairs whose typicality was difficult to classify (e.g., "good"–"job"; "little"–"bit"). To resolve this issue, we used human judgments of words' concreteness to identify and exclude candidate idioms (Brysbaert, Warriner, & Kuperman, 2014). We retained for analysis only pairs in which both the adjective and noun were in the top 25% of the concreteness ratings (e.g., "dirty" – "dish"; "green" – "fish") restricting our set to 2,477. Finally, human coders in the lab judged whether each pair was "incoherent or unrelated" and we excluded a final 576 pairs from the sample (e.g., incoherent pairs such as "flat" – "honey").

Thus, our final sample included 1,901 unique adjective-noun pairs drawn from 3,749 distinct utterances. The pairs were combinations of 637 distinct concrete nouns and 111 distinct concrete adjectives. We compiled these pairs and collected human judgments on Amazon Mechanical Turk for each pair, as described below. Table **??** contains example utterances from the final set and typicality judgments from our human raters.

Participants

Each participant rated 20 adjective-noun pairs, and each pair was rated by four participants; we used Dallinger, a tool for automating complex recruitment in online studies, to balance coding of the pairs. Overall, we recruited 444 participants to rate our final sample of adjective–noun pairs. After exclusions using an attention check that asked participants to simply choose a specific number on the scale, we retained 8,580 judgments, with each adjective–noun pair retaining at least two judgments.

Design and Procedure

To evaluate the typicality of the adjective–noun pairs that appeared in parents' speech, we asked participants on Amazon Mechanical Turk to rate each pair. Participants were presented with a question of the form "How common is it for a cow to be a brown cow?" and asked to provide a rating on a seven-point scale: (1) never, (2) rarely, (3) sometimes, (4) about half the time, (5) often, (6) almost always, (7) always.

Results

The human typicality ratings were combined with usage data from our corpus analysis to let us determine the extent to which parents use language to describe typical and atypical features. In our analyses, we token-weighted these judgments, giving higher weight to pairs that occurred more frequently in children's inputs. However, results are qualitatively identical and all significant effects remain significant without these re-weightings.

If caregivers speak informatively to convey what is atypical or surprising in relation to their own sophisticated world knowledge, we should see that caregiver description is dominated by adjectives that are sometimes or rarely true of the noun they modify. If instead child-directed speech privileges redundant information, perhaps to align to young children's limited world knowledge, caregiver description should yield a distinct distribution dominated by highly typical modifiers. As predicted in our pre-registration, we find that parents' description predominantly focuses on features that are atypical (Figure 1).

To confirm this effect statistically, we centered the ratings (i.e. "about half" was coded as 0), and then predicted the rating on each trial with a mixed effect model with only an intercept and a random effect of noun (`typicality ~ 1 + (1|noun)`). The intercept was reliably negative, indicating that adjectives tend to refer to atypical features of objects ($\beta =$ -0.77, $t =$ -19.72, $p < .001$). We then re-estimated these models separately for each age in the corpus, and found a reliably negative intercept for every age group (smallest effect $\beta_{14} =$ -0.50, $t =$ -4.45, $p = < .001$). These data suggest that even when talking with very young children, caregiver speech is structured according to the kind of communicative pressures observed in adult-adult conversation in the lab.

For comparison, we performed the same analyses but with typicality judgments weighted not by the frequency of each adjective-noun pair's occurrence in the Language Development Project, but instead by their frequency of occurrence in the Corpus of

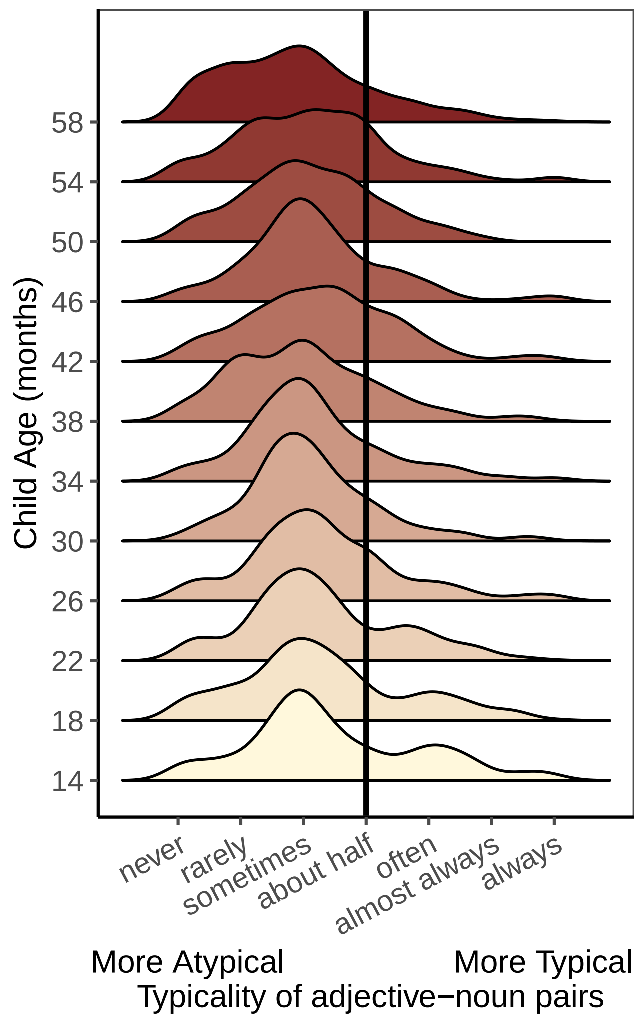

Figure 1. Density plots showing usage at each timepoint based on the typicality of the adjective-noun pair.

Contemporary American English (COCA; Davies, 2008). While this estimate of adult usage is imperfect—the adjective-nouns pairs produced by parents in our corpus may not be a representative sample of adjectives and nouns spoken by the adults in COCA—it provides a first approximation to adult usage. When we fit the same mixed-effects model to the data, we found that the intercept was reliably negative, indicating that adult-to-adult speech is likely also biased toward description of atypical features (β = -0.30, t = -19.72, p < .001). We propose to carry out a fuller analysis of adult-adult speech, using the same method we used for the LDP, for the full dissertation.

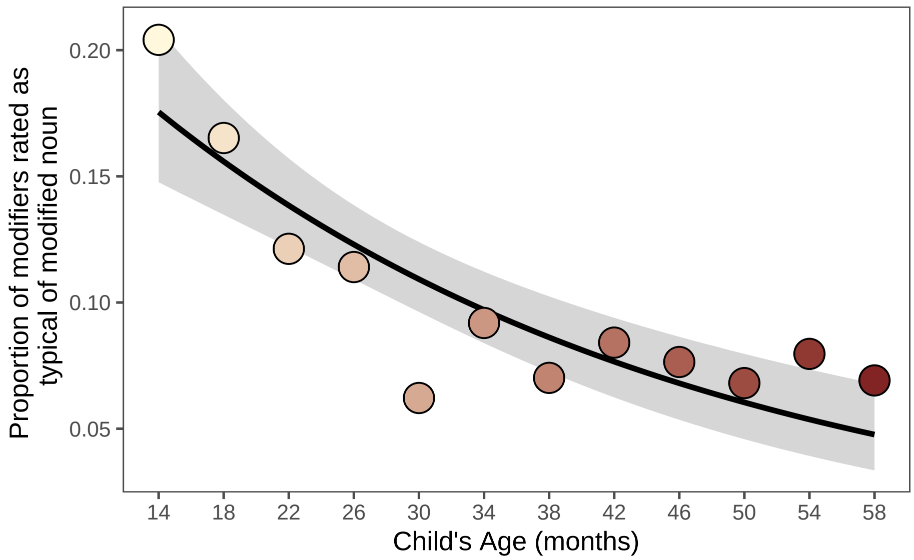

Figure 2. Proportion of caregiver description that is about typically-true features, as a function of age.

Returning to caregiver speech, while descriptions at every age tended to point out atypical features (as in adult-to-adult speech), this effect changed in strength over development. As predicted, an age effect added to the previous model was reliably negative, indicating that parents of older children are relatively more likely to focus on atypical features (β = -0.11, t = -3.47, p = .001). In line with the idea that caregivers adapt their speech to their children's knowledge, it seems that caregivers are more likely to provide description of typical features for their young children, compared with older children. As a second test of this idea, we defined adjectives as highly typical if Turkers judged them to be 'often', 'almost always', or 'always' true. We predicted whether each judgment was highly typical from a mixed-effects logistic regression with a fixed effect of age (log-scaled) and a random effect of noun. Age was a highly reliable predictor (β = -0.94, t = -5.01, p = $<$ .001). While children at all ages hear more talk about what is atypically true (Figure 1), younger children hear relatively more talk about what is typically true than older children do (Figure 2).

Discussion

In sum, we find robust evidence that language is used to discuss atypical, rather than typical, features of the world. Description in caregiver speech seems to largely mirror the usage patterns that we observed in adult-to-adult speech, suggesting that these patterns arise from general communicative pressures. Interestingly, the descriptions children hear change over development, becoming increasingly focused on atypical features. The higher prevalence of typical descriptors in early development may help young learners learn what is typical; however, even at the earliest point we measured, the bulk of language input describes atypical features.

This usage pattern aligns with the idea that language is used informatively in relation to background knowledge about the world. It may pose a problem, however, for young language learners with still-developing world knowledge. If language does not transparently convey the typical features of objects, and instead (perhaps misleadingly) notes the atypical ones, how might children come to learn what objects are typically like? One possibility is that information about typical features is captured in more complex regularities across many utterances. If this is true, language may still be an important source of information about typicality as children may be able to extract more accurate typicality information by tracking second-order co-occurrence.

Extracting Typicality from Language Structure

Much information can be gleaned from language that does not seem available at first glance. From language alone, simple distributional learning models can recover enough information to perform comparably to non-native college applicants on the Test of English as a Foreign Language (Landauer & Dumais, 1997). Recently, Lewis, Zettersten, and Lupyan (2019) demonstrated that even nuanced feature information may be learnable through distributional semantics alone, without any complex inferential machinery. We take a similar approach to ask whether a distributional semantics model trained on the language children hear can capture typical feature information.

Method

To test this possibility, we trained word2vec–a distributional semantics model–on the same corpus of child-directed speech used in our first set of analyses. Word2vec is a neural network model that learns to predict words from the contexts in which they appear. This leads word2vec to encode words that appear in similar contexts as similar to one another (Firth, 1957).

We used the continuous-bag-of-words (CBOW) implementation of word2vec in the `gensim` package (Řehůřek & Sojka, 2010). We trained the model using a surrounding context of 5 words on either side of the target word and 100 dimensions (weights in the hidden layer) to represent each word. After training, we extracted the hidden layer representation of each word in the model's vocabulary–these are the vectors used to represent these words.

If the model captures information about the typical features of objects, we should see that the model's noun-adjective word pair similarities are correlated with the typicality ratings we elicited from human raters. For a second comparison, we also used an off-the-shelf implementation of word2vec trained on Wikipedia (Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2018). While the Language Development Project corpus likely underestimates the amount of structure in children's linguistic input, Wikipedia likely overestimates it.

While word2vec straightforwardly represents what can be learned about word similarity by associating words with similar contexts, it does not represent the cutting edge of language modeling. Perhaps a more sophisticated model, trained on a larger corpus, would represent these typicalities better. To test this, we asked how BERT (Devlin, Chang, Lee, & Toutanova, 2018), a masked language model trained on English Wikipedia and BookCorpus, represents typicality. BERT does not directly provide similarity metrics between words, so to ask this, we must embed the pairs in sentential contexts. Since the placement of the adjective in the sentence may affect BERT's judgments, we used both a prenominal adjective sentence frame, which intuitively may express more atypical information, and a predicate adjective sentence frame, which intuitively may express more typical information. We gave BERT sentences of the form "I saw the ______ apple" (prenominal frame) and "The apple is ______" (predicate frame), and asked it the probability of different adjectives filling the empty slot. Because BERT has more complex training objectives and is trained on a much larger corpus than word2vec, results from BERT likely do not straightforwardly represent the information available to children in language. However, results from BERT can indicate the challenges language models face in representing world knowledge when the language people use emphasizes remarkable rather than typical situations.

Results

We find that similarities in the model trained on the Language Development Project corpus have near zero correlation with human adjective–noun typicality ratings ($r = 0.03$, $p = .208$). However, our model does capture other meaningful information about the structure of language, such as similarity. Comparing with pre-existing large-scale human similarity judgements for word pairs, our model shows significant correlations (correlation with wordsim353 similarities of noun pairs, 0.28; correlation with simlex similarities of noun, adjective, and verb pairs, 0.16). This suggests that statistical patterns in child-directed speech are likely insufficient to encode information about the typical features of objects, despite encoding at least some information about word meaning more broadly.

However, the corpus on which we trained this model was small; perhaps our model did not get enough language to draw out the patterns that would reflect the typical features of objects. To test this possibility, we asked whether word vectors trained on a much larger corpus—English Wikipedia—correlate with typicality ratings. This model's similarities were significantly correlated with human judgments, although the strength of the correlation was still fairly weak ($r = 0.25$, $p < .001$). How does an even larger and more sophisticated language model, BERT, fare? Like Wikipedia-trained word2vec, BERT's probabilities were significantly correlated with human judgments, though weakly so (prenominal adjective: $r = 0.22$, $p < .001$; predicate adjective: $r = 0.20$, $p < .001$).

One possible confound in these analyses is that the similarity judgments produced by our models reflect many dimensions of similarity, but our human judgments reflect only typicality. To accommodate this, we performed a second analysis in which we considered only the subset of 73 nouns that had both a typical (rated as at least "often") and an atypical (rated as at most "sometimes") adjective. We then asked whether the models rated the typical adjective as more similar to the noun it modified than the atypical adjective. The LDP model correctly classified 38 out of 73 (0.52), which was not better than chance ($p = .815$). The Wikipedia-trained word2vec model correctly classified 56 out of 73 (0.77), which was better than chance according to a binomial test, but still fairly poor performance ($p = < .001$). BERT correctly classified 42 out of 73 (0.58) in the prenominal sentence frame, which is not significantly better than chance ($p = .242$) and 49 (0.67) in the predicate sentence frame, which is significantly better than chance ($p = .005$). Both sets of BERT ratings are directionally less accurate than those of Wikipedia-trained word2vec: using a more advanced model did not improve performance on this task. Figure 3 shows the ratings from Turkers and the models for the 73 nouns and their typical and atypical adjectives. Table 1 gives the six cases in which word2vec similarities are worst at predicting human typicality judgments, judging the low-typicality adjective to be *more* similar to the noun than the high-typicality adjective.

General Discussion

Language provides children a rich source of information about the world. However, this information is not always transparently available: because language is used to comment on the atypical, it does not perfectly mirror the world. Among adult conversational partners whose world knowledge is well-aligned, this principle allows people to converse informatively

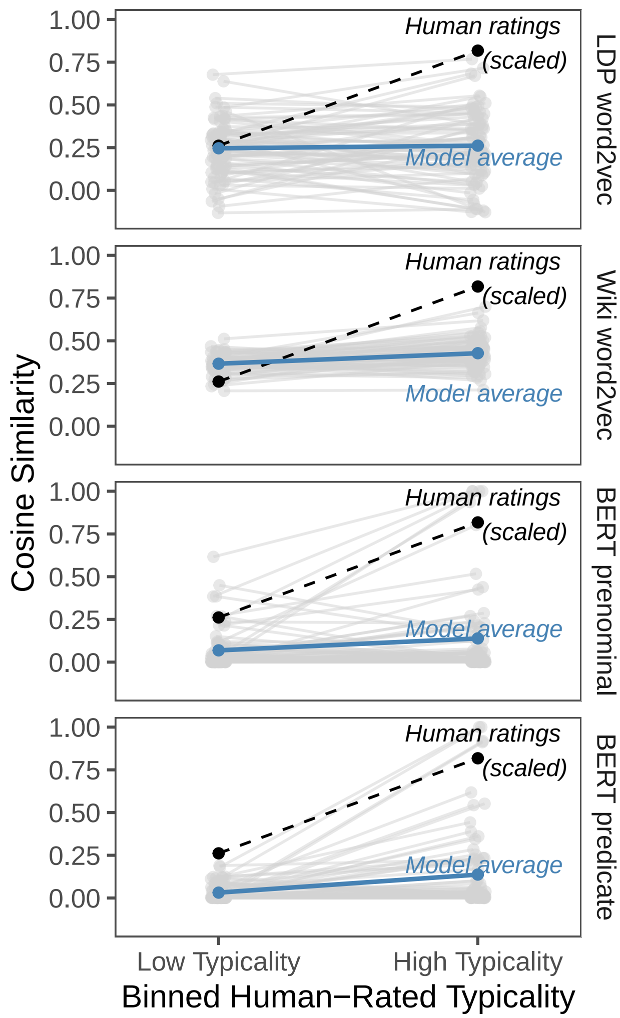

Figure 3. Plots of word2vec and BERT noun-adjective similarities for nouns for which there was at least one atypical adjective (rated at most "sometimes"), and at least one typical adjective (rated at least "often").

and avoid redundancy. But between a child and caregiver whose world knowledge is asymmetric, this pressure competes with other demands: what is minimally informative to an adult may be misleading to a child. Our results show that this pressure structures language to create a peculiar learning environment, one in which caregivers predominantly point out the atypical features of things.

How, then, do children learn about the typical features of things? While younger children may gain an important foothold from hearing more description of typical features,

noun	typical adjective	atypical adjective
puzzle	flat	giant
apple	red	brown
bird	outside	purple
elephant	fat	pink
whale	wet	red
frog	green	purple

Table 1

The top six cases in which Wikipedia-trained word2vec similarities were worst at predicting human typicality judgments. In each case, word2vec judged the low-typicality adjective to be more similar to the noun than the high-typicality adjective.

they still face language dominated by atypical description. When we looked at more nuanced ways of extracting information from language (which may or may not be available to the developing learner), we found that models of distributional semantics capture little typical feature information.

Of course, perceptual information from the world may simplify this problem. In many cases, perceptual information may swamp information from language; children likely see enough orange carrots in the world to outweigh hearing "purple carrot." It remains unclear, however, how children learn about categories for which they have scarcer evidence. Indeed, language information likely swamps perceptual information for many other categories, such as abstract concepts or those that cannot be learned about by direct experience. If such concepts pattern similarly to the concrete objects analyzed here, children are in a particularly difficult bind.

It is also possible that other cues from language and interaction provide young learners with clues to what is typical or atypical, and these cues are uncaptured by our measure of usage statistics. Caregivers may highlight when a feature is typical by using certain syntactic constructions, such as generics (e.g., “tomatoes are red”). Caregivers may also mark the atypicality of a feature, for example demonstrating surprise. Such cues from language and the interaction may provide key information in some cases; however, given the sheer frequency of atypical descriptors, it seems unlikely that they are consistently well-marked.

Another possibility is that children expect language to be used informatively at a young age. Under this hypothesis, their language environment is not misleading at all, even without additional cues from caregivers. Children as young as two years old tend to use words to comment on what is new rather than what is known or assumed (Baker & Greenfield, 1988). Children may therefore expect adjectives to comment on surprising features of objects. If young children expect adjectives to mark atypical features (Horowitz & Frank, 2016), they can use description and the lack thereof to learn more about the world. We propose to investigate this question in Chapter 3.

Across our analyses, language is used with remarkable consistency: people talk about the atypical. Though parents might reasonably be broadly over-informative in order to teach their children about the world, this is not the case. This presents a potential puzzle for young learners who have limited world knowledge and limited pragmatic inferential abilities. Perceptual information and nascent pragmatic abilities may help fill in the gaps, but much remains to be explored to link these explanations to actual learning. Communication pressures are pervasive forces structuring the language children hear, and future work must disentangle whether children capitalize on them or are misled by them in learning about the world.

Chapter 2: How adults use contrastive inference to learn about new categories

When referring to a *big red dog* or a *hot-air balloon*, we often take care to describe them—even when there are no other dogs or balloons around. Speakers use more description when referring to objects with atypical features (e.g., a yellow tomato) than typical ones (e.g., a red tomato; see Chapter 1 and Bergey, Morris, & Yurovsky, 2020; Mitchell et al., 2013; Rubio-Fernández, 2016a; Westerbeek et al., 2015a). This selective marking of atypical objects potentially supplies useful information to listeners: they have the opportunity to not only learn about the object at hand, but also about its broader category. Horowitz and Frank (2016) demonstrated that, combined with other contrastive cues (e.g., "Wow, this one is a zib. This one is a TALL zib"), prenominal adjectives prompted adults and children to infer that the described referent was less typical than one that differed on the mentioned feature (e.g., a shorter zib). In Chapter 2, we test whether listeners use descriptive contrast with a novel object's category to learn about the category's feature distribution.

If listeners do make contrastive inferences about typicality, it may not be as simple as judging that a described referent is atypical. Description can serve many purposes. If a descriptor is needed to distinguish between two present objects, it may not have been used to mark atypicality. For instance, in the context of a bin of heirloom tomatoes, a speaker who wants a red one in particular might specify that they want a "red tomato" rather than just asking for a "tomato." In this case, the adjective "red" is being used contrastively with respect to reference, and not to mark atypicality. If reference is the primary motivator of speakers' word choice, as implicitly assumed in much research (e.g., Pechmann, 1989; Arts, Maes, Noordman, & Jansen, 2011; Engelhardt et al., 2011), then people should draw no further inferences once the need for referential disambiguation explains away a descriptor like "red." On this reference-first view, establishing reference has priority in understanding the utterance, and any further inferences are blocked if the utterance is minimally informative with respect to reference. If, on the other hand, pragmatic reasoning weighs multiple goals simultaneously–here, reference and conveying typicality–people may integrate typicality as just one factor the speaker considers in using description, leading to graded inferences about the referent's identity and about its category's features.

In two experiments, we used an artificial language task to set up just this kind of learning situation. We manipulated the contexts in which listeners hear adjectives modifying novel names of novel referents. These contexts varied in how useful the adjective was to identify the referent: some contexts the adjectives were necessary for reference, and in others they were unhelpful. On a reference-first view, use of an adjective that was necessary for reference can be explained away and should not prompt further inferences about typicality–an atypicality inference would be blocked. If, on the other hand, people take into account speakers' multiple reasons for using adjectives without giving priority to reference, they may alter their inferences about typicality across these contexts in a graded way: if an adjective was necessary for reference, it may prompt slightly weaker inferences of atypicality; if an adjective was redundant with respect to reference, it may be inferred to mark atypicality more strongly. Further, these contexts may also prompt distinct inferences when no adjective is used: for instance, when an adjective is necessary to identify the referent but elided, people may infer that the elided feature is particularly typical. To account for the multiple ways context effects might emerge, we analyze both of these possibilities. Overall, we asked whether listeners infer that these adjectives identify atypical features of the named objects, and whether the strength of this inference depends on the referential ambiguity of the context in which adjectives are used.

Experiment 1

Method

Participants. 240 participants were recruited from Amazon Mechanical Turk. Half of the participants were assigned to a condition in which the critical feature was color (red, blue, purple, or green), and the other half of participants were assigned to a condition in which the critical feature was size (small or big).

Stimuli & Procedure. Stimulus displays showed two alien interlocutors, one on the left side (Alien A) and one on the right side (Alien B) of the screen, each with two novel

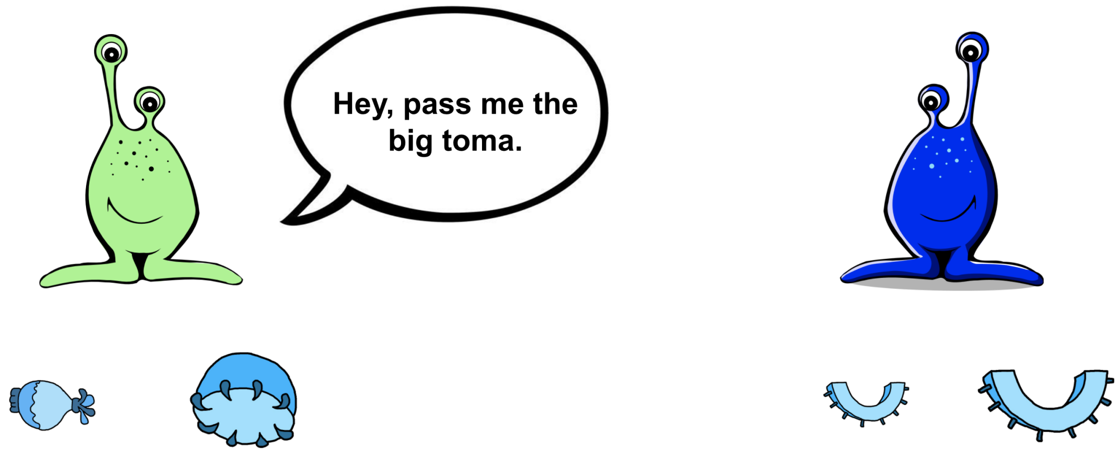

Figure 4. Experiment 1 stimuli. In the above example, the critical feature is size and the object context is a within-category contrast: the alien on the right has two same-shaped objects that differ in size.

fruit objects beneath them (Figure 4). Alien A, in a speech bubble, asked Alien B for one of its fruits (e.g., "Hey, pass me the big toma"). Alien B replied, "Here you go!" and the referent disappeared from Alien B's side and reappeared on Alien A's side.

We manipulated the critical feature type (color or size) between subjects. Two factors (presence of the critical adjective in the referring expression and object context) were fully crossed within subjects. Object context had three levels: within-category contrast, between-category contrast, and same feature (Figure 5). In the within-category contrast condition, Alien B possessed the target object and another object of the same shape, but with a different value of the critical feature (e.g., a big toma and a small toma). In the between-category contrast condition, Alien B possessed the target object and another object of a different shape, and with a different value of the critical feature (e.g., a big toma and a small blicket). In the same feature condition, Alien B possessed the target object and another object of a different shape but with the same value of the critical feature as the target (e.g., a big toma and a big dax). Thus, in the within-category contrast condition, the descriptor was necessary to distinguish the referent; in the between-category contrast condition it was unnecessary but potentially helpful; and in the same feature condition it was unnecessary and unhelpful.

Note that in all context conditions, the set of objects on screen was the same in terms of the experiment design: there was a target (e.g., big toma), an object with the same shape as the target and a different critical feature (e.g., small toma), an object with a different shape from the target and the same critical feature (e.g., big dax), and an object with a different shape from the target and a different critical feature (e.g., small blicket). Context was manipulated by rearranging these objects such that the relevant referents (the objects under Alien B) differed and the remaining objects were under Alien A. Thus, in each case, participants saw the target object and one other object that shared the target object's shape but not its critical feature–they observed the same kind of feature distribution of the target object's category in each trial type.

The particular values of the features were chosen randomly for each trial, and fruits were chosen randomly at each trial from 25 fruit kinds. Ten of the 25 fruit drawings were adapted and redrawn from Kanwisher, Woods, Iacoboni, and Mazziotta (1997); we designed the remaining 15 fruit kinds. Each fruit kind had an instance in each of four colors (red, blue, green, or purple) and two sizes (big or small).

Participants completed six trials. After each exchange between the alien interlocutors, they made a judgment about the prevalence of the target's critical feature in the target object's category. For instance, after seeing a red blicket being exchanged, participants would be asked, "On this planet, what percentage of blickets do you think are red?" They answered on a sliding scale between zero and 100. In the size condition, participants were asked, "On this planet, what percentage of blickets do you think are the size shown below?" with an image of the target object they just saw available on the screen.

After completing the study, participants were asked to select which of a set of alien words they had seen previously during the study. Four were words they had seen, and four were novel lure words. Participants were dropped from further analysis if they did not respond to at least 6 of these 8 correctly (above chance performance as indicated by a one-tailed binomial test at the $p = .05$ level). This resulted in excluding 47 participants, leaving 193 for further analysis.

Results

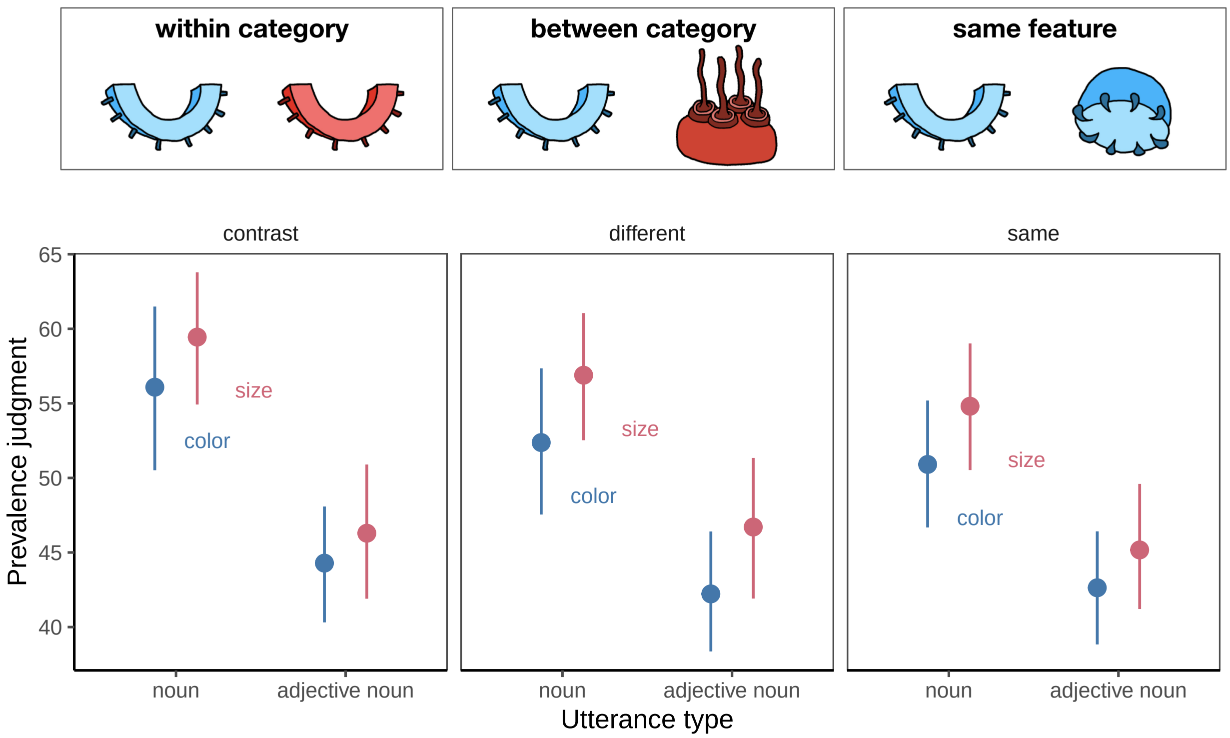

Figure 5. Prevalence judgments from Experiment 1. Participants consistently judged the target object as less typical of its category when the referent was described with an adjective (e.g., "Pass me the blue toma") than when it was not (e.g., "Pass me the toma"). This inference was not significantly modulated by object context (examples shown above each figure panel).

Our key test is whether participants infer that a mentioned feature is less typical than one that is not mentioned. In addition, we tested whether inferences of atypicality are modulated by context. One way to test this is to analyze the interaction between utterance type and context, seeing if the difference between adjective and no adjective utterances is larger when the adjective was highly redundant or smaller when the adjective was necessary for reference.

We analyzed participants' judgments of the prevalence of the target object's critical feature in its category. We began by fitting a maximum mixed-effects linear model with effects of utterance type (adjective or no adjective), context type (within category, between category, or same feature, with between category as the reference level), and critical feature (color or size) as well as all interactions and random slopes of utterance type and context type nested within subject. Random effects were removed until the model converged. The final model included the effects of utterance type, context type, and critical feature and their interactions, and a random slope of utterance type by subject.

This model revealed a significant effect of utterance type ($\beta_{adjective}$ = -10.22, t = -3.37, p = .001), such that prevalence judgments were lower when an adjective was used than when it was not. Participants' inferences did not significantly differ between color and size adjective conditions (β_{size} = 4.73, t = 1.46, p = .146). Participants' inferences did not significantly vary by context type (β_{within} = 3.92, t = 1.63, p = .104; β_{same} = -1.48, t = -0.62, p = .537). There was not a significant interaction between context and presence of an adjective in the utterance ($\beta_{within*adjective}$ = -1.58, t = -0.46, p = .644; $\beta_{same*adjective}$ = 2.13, t = 0.63, p = .532). That is, participants did not significantly adjust their inferences based on object context, nor did they make differential inferences based on the combination of context and adjective use. However, they robustly inferred that mentioned features were less prevalent in the target's category than unmentioned features.

This lack of a context effect may be because people do not take context into account, or because they make distinct inferences when an adjective is *not* used: for instance, when an adjective is necessary for reference but elided, people may infer that the unmentioned feature is very typical. This inference would lead to a difference between the adjective and no adjective utterances in the within-category context, but not because people are failing to attribute the adjective to reference. To account for this possibility, we additionally tested for differences in the context conditions among only the utterances with adjectives. We fit a model with effects of context type and critical feature as well as their interaction and random slopes by subject. Participants did not significantly adjust their inferences by context among only the adjective utterances ($\beta_{within} = 2.43$, $t = 1.16$, $p = .247$; $\beta_{same} = 0.67$, $t = 0.32$, $p = .750$). Thus, even by this more specific test, participants did not adjust their inferences based on the referential context.

Discussion

Description is often used not to distinguish among present objects, but to pick out an object's feature as atypical of its category. In Experiment 1, we asked whether people would infer that a described feature is atypical of a novel category after hearing it mentioned in an exchange. We found that people robustly inferred that a mentioned feature was atypical of its category, across both size and color description. Further, participants did not use object context to substantially explain away description. That is, even when description was necessary to distinguish among present objects (e.g., there were two same-shaped objects that differed only in the mentioned feature), participants still inferred that the feature was atypical of its category. This suggests that, in the case of hearing someone ask for a "red tomato" from a bin of many-colored heirloom tomatoes, a tomato-naive person would infer that tomatoes are relatively unlikely to be red.

Model

To formalize the inference that participants were asked to make, we developed a model in the Rational Speech Act Framework (RSA, Frank & Goodman, 2012). In this framework, pragmatic listeners (L) are modeled as drawing inferences about speakers' (S) communicative intentions in talking to a hypothetical literal listener (L_0). This literal listener makes no pragmatic inferences at all, evaluating the literal truth of a statement (e.g., it is true that a red toma can be called “toma” and “red toma” but not “blue toma”), and chooses randomly among all referents consistent with that statement. In planning their referring expressions, speakers choose utterances that are successful at accomplishing two goals: (1) making the listener as likely as possible to select the correct object, and (2) minimizing their communicative cost (i.e., producing as few words as possible). Note that though determiners are not given in the model’s utterances, the assumption that the utterance refers to a specific reference is built into the model structure, consistent with the definite determiners used in the task. Pragmatic listeners use Bayes’ rule to invert the speaker’s utility function, essentially inferring what the speaker’s intention was likely to be given the utterance they produced.

$$Literal : P_{Lit} = \delta\left(u, r\right) P\left(r\right)$$

$$Speaker : P_S\left(u|r\right) \propto \alpha\left(P_{Lit}\left(r|u\right) - C\right)$$

$$Listener : P_{Learn}\left(r|u\right) \propto P_s\left(u|r\right) P\left(r\right)$$

To allow the Rational Speech Act Framework to capture inferences about typicality, we modified the Speaker’s utility function to have an additional term: the listener’s expected processing difficulty. Speakers may be motivated to help listeners to select the correct referent not just eventually but as quickly as possible. People are both slower and less accurate at identifying atypical members of a category as members of that category (Dale, Kehoe, & Spivey, 2007; Rosch, Simpson, & Miller, 1976). If speakers account for listeners’ processing difficulties, they should be unlikely to produce bare nouns to refer to low typicality exemplars (e.g. unlikely to call a purple carrot simply “carrot”). This is roughly the kind of inference encoded in a continuous semantics Rational Speech Act model (Degen, Hawkins, Graf, Kreiss, & Goodman, 2020).

We model the speaker as reasoning about the listener's label verification process. Because the speed of verification scales with the typicality of a referent, a natural way of modeling it is as a process of searching for that particular referent in the set of all exemplars of the named category, or alternatively of sampling that particular referent from the set of all exemplars in that category, $P(r|Cat)$. On this account, speakers want to provide a modifying adjective for atypical referents because the probability of sampling them from their category is low, but the probability of sampling them from the modified category is much higher (a generalization of the size principle, Xu & Tenenbaum, 2007). Typicality is just one term in the speaker's utility, and thus is directly weighed with the literal listener's judgment and against cost.

If speakers use this utility function, a listener who does not know the feature distribution for a category can use a speaker's utterance to infer it. Intuitively, a speaker should prefer not to modify nouns with adjectives because they incur a cost for producing an extra word. If they did use an adjective, it must be because they thought the learner would have a difficult time finding the referent from a bare noun alone because of typicality, competing referents, or both. To infer the true prevalence of the target feature in the category, learners combine the speaker's utterance with their prior beliefs about the feature distribution. We model the learner's prior about the prevalance of features in any category as a Beta distribution with two parameters α and β that encode the number of hypothesized prior psuedo-exemplars with the feature and without feature that the learner has previously observed (e.g., one red dax and one blue dax). We assume that the learner believes they have previously observed one hypothetical psuedo-examplar of each type, which is a weak symmetric prior indicating that the learner expects the target feature value to occur in half of all members of a category on average, but would find many levels of prevalence unsurprising. To model the learner's direct experience with the category, we add the observed instances in the experiment to these hypothesized prior instances. After observing one member of the category with the target feature value and one without, the listener's prior is thus updated to be Beta $(2, 2)$.

We used Bayesian data analysis to estimate the posterior mean rationality parameter that participants are using to draw inferences about speakers in both the color and size conditions. The absolute values of these parameters are driven largely by the number of pseudo-exemplars assumed by the listener prior to exposure; however, differences between color and size within the model are interpretable. We found that listeners inferred speakers to be directionally more rational when using size adjectives (0.89 [0.63, 1.13]) than color adjectives (0.60 [0.37, 0.83]), but the two inferred confidence intervals were overlapping, suggesting that people treated size and color adjectives similarly when making inferences about typicality.

Figure 6 shows the predictions of our Rational Speech Act model compared to empirical data from participants. The model captures the trends in the data correctly, inferring that the critical feature was less prevalent in the category when it was mentioned (e.g., "red dax") than when it was not mentioned (e.g., "dax"). The model also infers the prevalence of the critical feature to be numerically higher in the within-category condition, like people do. That is, in the within-category condition when an adjective is used to distinguish between referents, the model thinks that the target color is slightly less atypical. When an adjective would be useful to distinguish between two objects of the same shape but one is not used, the model infers that the color of the target object is slightly more typical.

Overall, our model captures the inference people make: when the speaker mentions a feature (e.g., "the blue dax"), people infer that the feature is less typical of the category (daxes are less likely to be blue in general). It further captures that when the object context requires an adjective for successful reference, people weaken this atypicality inference only

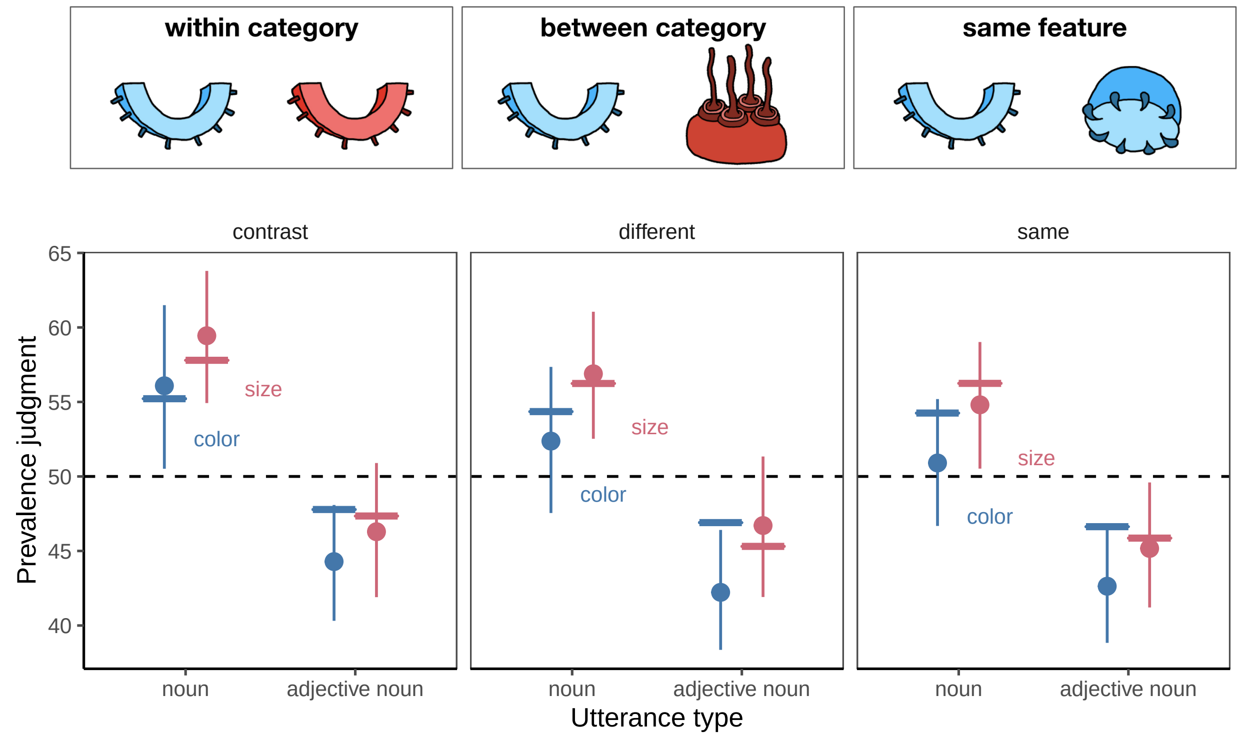

Figure 6. Participants' prevalence judgments from Experiment 1, compared to model predictions (horizontal lines).

slightly, if at all. In contrast to a reference-first view, which predicts that these two kinds of inferences would trade off strongly–that is, using an adjective that is necessary for reference would block the inference that it is marking atypicality–the model captures the graded way in which people consider these two communicative goals.

Experiment 2

In Experiments 1, we established that people can use contrastive inferences to make inferences about the feature distribution of a novel category. Additionally, we found that these two inferences do not seem to trade off substantially: even if an adjective is necessary to establish reference, people infer that it also marks atypicality. To strengthen our findings in a way that would allow us to better detect potential trade-offs between these two types of inference, in Experiment 2 we conducted a pre-registered replication of Experiment 1 with a larger sample of participants. In addition, we tested how people's prevalence judgments from utterances with and without an adjective compare to their null inference about feature prevalence by adding a control utterance condition: an alien utterance, which the participants could not understand. This also tests the model assumption we made in Experiment 1: that after seeing two exemplars of the target object with two values of the feature (e.g., one green and one blue), people's prevalence judgments would be around 50%. In addition to validating this model assumption, we more strongly tested the model here by comparing predictions from same model, with parameters inferred from the Experiment 1 data, to data from Experiment 2. Our pre-registration of the method, recruitment plan, exclusion criteria, and analyses can be found on the Open Science Framework: https://osf.io/s8gre .

Method

Participants. A pre-registered sample of 400 participants was recruited from Amazon Mechanical Turk. Half of the participants were assigned to a condition in which the critical feature was color (red, blue, purple, or green), and half of the participants were assigned to a condition in which the critical feature was size (small or big).

Stimuli & Procedure. The stimuli and procedure were identical to those of Experiment 2, with the following modifications. Two factors, utterance type and object context, were fully crossed within subjects. Object context had two levels: within-category contrast and between-category contrast. In the within-category context condition, Alien B possessed the target object and another object of the same shape, but with a different value of the critical feature (color or size). In the between-category contrast condition, Alien B possessed the target object and another object of a different shape, and with a different value of the critical feature. Thus, in the within-category contrast condition, an adjective is necessary to distinguish the referent; in the between-category contrast condition it is unnecessary but potentially helpful. There were three utterance types: adjective, no adjective, and alien utterance. In the two alien utterance trials, the aliens spoke using completely unfamiliar utterances (e.g., "Zem, noba bi yix blicket"). Participants were told in the task instructions that sometimes the aliens would talk in a completely alien language, and sometimes their language will be partly translated into English. To keep participants from making inferences about the content of the alien utterances using the utterance content of other trials, both alien language trials were first; other than this constraint, trial order was random. We manipulated the critical feature type (color or size) between subjects.

After completing the study, participants were asked to select which of a set of alien words they had seen previously during the study. Four were words they had seen, and four were novel lure words. Participants were dropped from further analysis if they did not meet our pre-registered criteria of responding to at least 6 of these 8 correctly (above chance performance as indicated by a one-tailed binomial test at the $p = .05$ level) and answering all four color perception check questions correctly. Additionally, six participants were excluded because their trial conditions were not balanced due to an error in the run of the experiment. This resulted in excluding 203 participants, leaving 197 for further analysis. In our pre-registration, we noted that we anticipated high exclusion rates, estimating that approximately 150 people per condition would be sufficient to test our hypotheses.

Results

We began by fitting a pre-registered maximum mixed-effects linear model with effects of utterance type (alien utterance, adjective, or no adjective; alien utterance as reference level), context type (within category or between category), and critical feature (color or size) as well as all interactions and random slopes of utterance type and context type nested within subject. Random effects were removed until the model converged, which resulted in a model with all fixed effects, all interactions and a random slope of utterance type by subject. The final model revealed a significant effect of the no adjective utterance type compared to the alien utterance type ($\beta = 7.48$, $t = 2.80$, $p = .005$) and no significant effect of the adjective utterance type compared to the alien utterance type ($\beta =$ -0.64, $t =$ -0.24, $p =$

.808). The effects of context type (within-category or between-category) and adjective type (color or size) were not significant (β_{within} = -2.70, t_{within} = -1.23, p_{within} = .220; β_{size} = 4.44, t_{size} = 1.33, p_{size} = .185). There were marginal interactions between the adjective utterance type and the size condition (β = -6.56, t = -1.72, p = .086), the adjective utterance type and the within-category context (β = 5.77, t = 1.86, p = .064), and the no adjective utterance type and the within-category context (β = 5.57, t = 1.79, p = .073). No other effects were significant or marginally significant. Thus, participants inferred that an object referred to in an intelligible utterance with no description was more typical of its category on the target feature than an object referred to with an alien utterance. Participants did not substantially adjust their inferences based on the object context. The marginal interactions between the within-category context and both the adjective and no adjective utterance types suggest that people might have judged the target feature as slightly more prevalent in the within-category context when intelligible utterances (with a bare noun or with an adjective) were used compared to the alien utterance. If people are discounting their atypicality inferences when the adjective is necessary for reference, we should expect them to have slightly higher typicality judgments in the within-category context when an adjective is used, and this marginal interaction suggests that this may be the case. However, since typicality judgments in the no adjective utterance type are also marginally greater in the within-category context, and because judgments in the alien utterance conditions (the reference category) also directionally move between the two context conditions, it is hard to interpret whether this interaction supports the idea that people are discounting their typicality judgments based on context.

Given that interpretation of these results with respect to the alien utterance condition can be difficult, we pre-registered a version of the same full model excluding alien utterance trials with the no adjective utterance type as the reference level. This model revealed a significant effect of utterance type: participants' prevalence judgments were lower when an adjective was used than when it was not (β = -8.12, t = -3.46, p = .001). No other effects were significant. This replicates the main effect of interest in Experiment 1: when an adjective is used in referring to the object, participants infer that the described feature is less typical of that object's category than when the feature goes unmentioned. It also shows that the possibility that people may discount their typicality judgments based on context (suggested by the marginal interaction described above) is not supported when we compare the adjective and no adjective utterance types directly. In the Supplemental Materials, we report two more pre-registered tests of the effect of utterance type alone on prevalence judgments whose results are consistent with the fuller models reported here.

As in Experiment 1, our test of whether participants' inferences are modulated by context is potentially complicated by people making distinct inferences when an adjective is necessary but *not* used. Thus, we additionally tested whether participants' inferences varied by context among only utterances with an adjective by fitting a model with effects of context and adjective type and their interaction, as well as random slopes by subject (not pre-registered). Participants' inferences did not significantly differ by context ($\beta_{within} = 3.07$, $t_{within} = 1.70$, $p_{within} = .091$). Thus, participants' inferences did not significantly differ between contexts, whether tested by the interaction between utterance type and contexts or by the effect of context among only utterances with an adjective.

Model

To validate the model we developed for Experiment 1, we compared its estimates using the previously fit parameters to the new data from Experiment 2. As shown in Figure 7, the model predictions were well aligned with people's prevalence judgments. In addition, in Experiment 1, we fixed the model's prior beliefs about the prevalence of the target object's color or size to be centered at 50% because the model had seen one pseudo-exemplar of the target color/size, and one psuedo-exemplar of the non-target color/size. In Experiment 2, we aimed to estimate this prior empirically in the alien utterance condition, reasoning that people could only use their prior to make a prevalence judgment (as we asked the model to

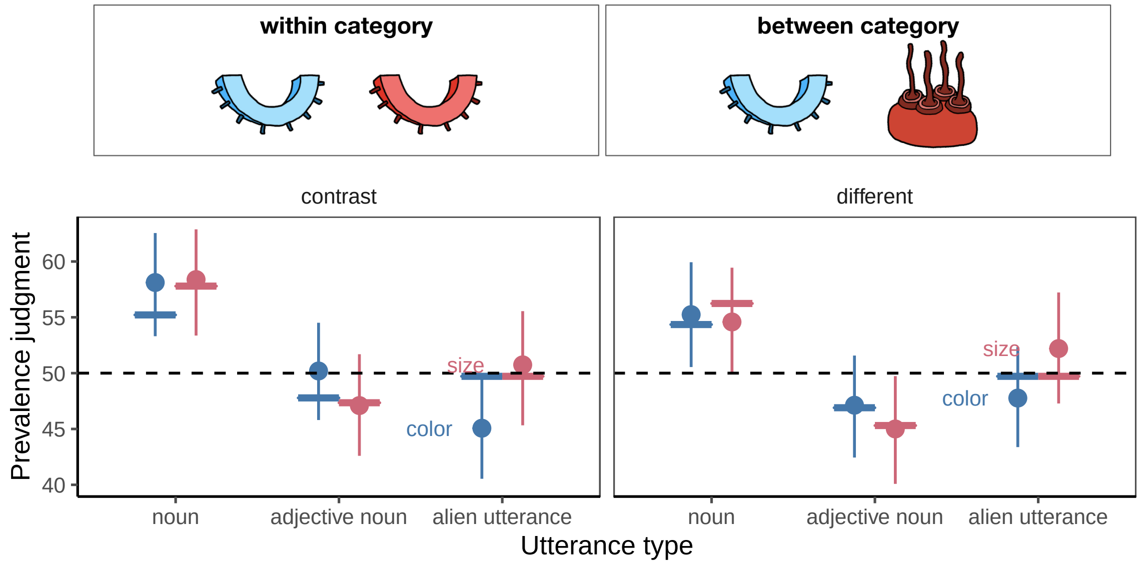

Figure 7. Participants' prevalence judgments in Experiment 2, with model predictions using the parameters estimated in Experiment 1 (horizontal lines).

do). In both the color and size conditions, people's judgments indeed varied around 50%, although in the color condition they were directionally lower. This small effect may arise from the fact that size varies on a scale with fewer nameable points (e.g., objects can be big, medium-sized or small) whereas color has many nameable alternatives (e.g., red, blue, green, etc.). Thus, the results of Experiment 2 confirm the modeling assumptions we made in estimating people's prior beliefs, and further validate the model we developed as a good candidate model for how people simultaneously draw inferences about speakers' intended referents and the typicality of these referents. That is, when people think about why a speaker chose their referring expression, they consider the context of not only present objects, but also the broader category to which the referent belongs.

Discussion

In Experiment 2, we replicated the main finding of interest in Experiment 1: when a novel object's feature is described, people infer that the feature is rarer of its category than when it goes unmentioned. Again, this effect was consistent across both size and color adjectives, and people did not substantially adjust this inference based on how necessary the description was to distinguish among potential referents. We also added an alien language condition, in which the entire referring expression was unintelligible to participants, to probe people's priors on feature typicality. We found that in the alien language condition, people judged features to be roughly between the adjective utterance and no adjective utterance conditions, and significantly different from the no adjective utterance condition. In the alien language condition, people's prevalence judgments were roughly around our model's prevalence judgments (50%) after observing the objects on each trial and before any inferences about the utterance.

The similarity of people's prevalence judgments in the alien language condition and the adjective condition raises the question: is this effect driven by an atypicality inference in the adjective conditions, or a *typicality* inference when the feature is unmentioned? Our results suggest that it is a bit of both. When someone mentions an object without extra description, the listener can infer that its features are likely more typical than their prior; when they use description, they can infer that its features are likely less typical. Because using an extra word–an adjective–is generally not thought of as the default way to refer to something, this effect is still best described as a contrastive inference of *atypicality* when people use description. However, the fact that people infer high typicality when an object is referred to without description suggests that, in some sense, there is no neutral way to refer: people will make broader inferences about a category from even simple mentions of an object.

Conclusion

In Chapter 1, we established that people tend to mention atypical rather than typical features. In this chapter, we showed that adults make appropriate pragmatic inferences given how speakers describe: they infer that a mentioned feature is likely to be less typical of the mentioned category. However, the ability to learn about new categories using contrastive inference most obviously serves budding language learners—children. To fully appreciate the potential of these inferences to allow people to learn about the world, we must study their development, which we will turn to in Chapter 3.

Chapter 3: How children use contrastive inference to learn about new categories

The speech children hear mentions more atypical than typical features. Depending on children's pragmatic abilities, this input could provide helpful information or pose a misleading challenge as children learn about the world. If children are able to make the contrastive inference that description tends to pick out atypical features, they could use description to go beyond learning about what they directly experience. If, on the other hand, they merely associate the mentioned feature with the mentioned category, they may mistakenly learn that atypical features are more common than they actually are.

In general, children's pragmatic abilities are thought to undergo prolonged development, not reaching adult-like performance until well into schooling age. The most thoroughly studied pragmatic inference in children, scalar implicature, tells a bleak story about children's ability to make pragmatic inferences at a young age. Scalar implicature is the phenomenon in which use of a weak scalar term ('some,' 'might') implies that a stronger scalar term ('all,' 'must') is not true—for example, "I ate some of the cookies" implies I did not eat all of them. This inference can be derived by reasoning that had the speaker meant the stronger meaning, they would have used the stronger term. Adults consistently interpret the word 'some' to mean 'some but not all,' rating the use of 'some' as unnatural when 'all' is applicable and taking longer to respond to such instances (Bott & Noveck, 2004; Degen & Tanenhaus, 2015). Until at least the age of 5 and in some tasks up to 10 years old, children fail to limit the use of 'some' in this way, accepting 'some' as a descriptor when 'all' is true (Noveck, 2001; Papafragou & Musolino, 2003). This deficit is found in a range of measures, from acceptability judgments to eye-tracking (Huang & Snedeker, 2009). Later work has found that children likely lack this ability because they fail to activate alternative descriptions, so cannot reason that the speaker should have said ‘all’ and not ‘some’ if all is true (Barner, Brooks, & Bale, 2011), and because they lack a meta-understanding of these tasks (Papafragou & Musolino, 2003). When given supportive context, like named alternatives or training on the task, 4- and 5-year-olds improve at these implicatures (Barner et al., 2011; Foppolo, Guasti, & Chierchia, 2012; Papafragou & Musolino, 2003). However, across experiments, performance on scalar implicature remains fragile well into school age.

Contrastive inference from description, however, may be a more accessible form of pragmatic inference because the relevant alternatives are more easily accessible. In the case of using contrastive inference to resolve reference (e.g., “the tall. . . ” prompts looking to a tall object with a shorter counterpart), the relevant alternatives are available in the environment. By the age of 5, children can use contrastive inferences to direct their attention among familiar present objects (Huang & Snedeker, 2008), and when given extra time to orient to the referent, show budding abilities by the age of 3 (Davies, Lingwood, Ivanova, & Arunachalam, 2021). Description paired with other contrastive cues can allow children to restrict reference among novel objects or objects with novel properties, though imperfectly (Diesendruck, Hall, & Graham, 2006; Gelman & Markman, 1985).

What about when the contrasting set is not available in the environment, but is the referent’s category? Preliminary evidence also suggests that contrastive inference about typicality may be possible for young children. When paired with other contrastive cues, 4-year-olds can make inferences about novel object typicality, reasoning that “the TALL zib” suggests other zibs are generally shorter (Horowitz & Frank, 2016). This work provided a useful demonstration that adjective use can contribute to inferences about feature typicality, though it did not isolate the effect of adjectives specifically. Their experiments used several contrastive cues, such as prosody (contrastive stress on the adjective: “TALL zib”), demonstrative phrases that may have marked the object as unique (“this one”) and expressions of surprise at the object (“wow”), and participants may have inferred the object was atypical primarily from these cues and not from the adjective. Further, these experiments used a forced-choice measure that does not allow a precise estimate of how much children's typicality judgments shift from adjective use. Thus, in this experiment, we set out to develop a task that would isolate the effect of adjective use and measure children's typicality judgments in a more graded way.

In this chapter, we report an exploratory study of children's abilities to make contrastive inferences about typicality. To do this, we used a task similar to those done by adults in Chapter 2, having children observe novel categories and make inferences about the typicality of their features. We study 5- to 6-year-old children, an age at which key pragmatic abilities are developing and when children can use contrastive inferences to direct their attention among familiar referents. Because children at this age struggle to explicitly reason about and report proportions (see Boyer, Levine, & Huttenlocher, 2008 for a review), we will have children report their typicality judgments with the help of visual depictions of *few*, *some*, *most*, and *almost all* objects having a feature. The purpose of this exploratory study is both to see whether children can make sensible responses on this measure and to gather preliminary evidence about children's contrastive inferences.

Method

Participants. We recruited 29 5–6-year-old children raised with 90% or greater English language exposure to participate in this task. Children were recruited from a database with mostly families living in the Chicago area, and some families living elsewhere in the United States. In the final sample, 15 5-year-olds and 14 6-year-olds participated.

Design and Procedure. Children participated in a novel object learning task in which they observed novel objects and made inferences about them. They were introduced to an alien named Blip, who would show things from her planet. Blip's utterances were presented both as recorded audio and displayed in a text bubble on the screen. In each trial, Blip first said "Let's see what I have. . . " and then sequentially showed two objects with the

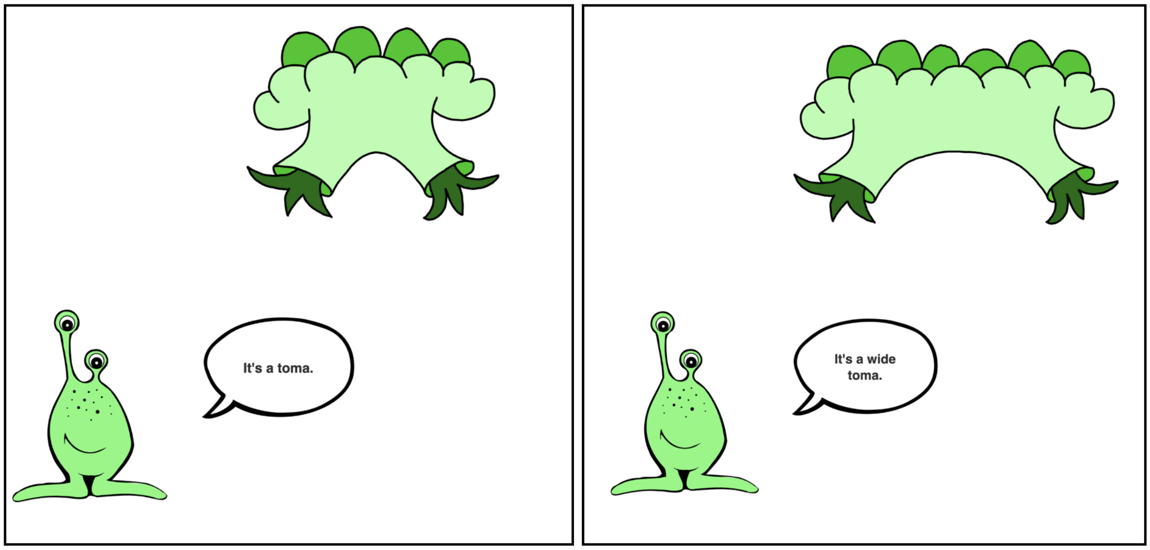

Figure 8. An example of the novel objects shown on a trial. In each trial, two objects of the same shape and differing on the critical feature were shown sequentially. In adjective noun trials, the critical feature was mentioned for the object that had it (e.g., the wide blicket was called a "wide blicket") and in noun trials, no features were mentioned (e.g., both blickets were just called a "blicket".

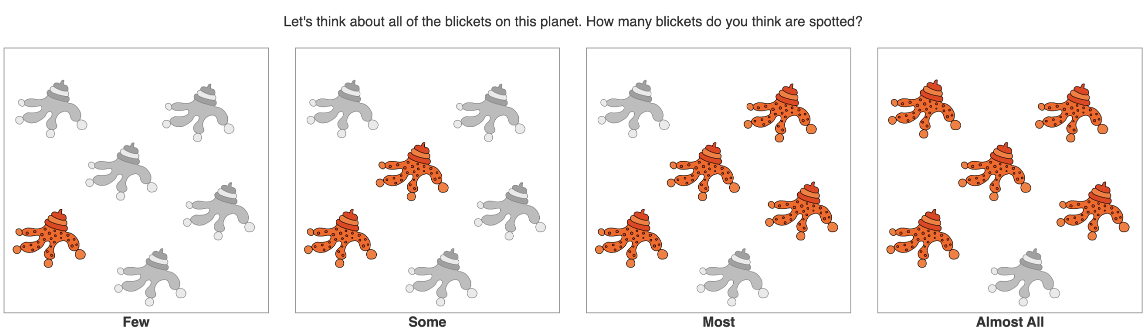

Figure 9. An example of the prevalence judgment children were asked to make. Children chose between clouds of novel objects representing few, some, most, and almost all of the novel category having the feature. The experimenter asked, e.g., "Let's think about all of the tomas on this planet. How many tomas do you think are spotted? Few of the tomas, some of the tomas, most of the tomas, or almost all of the tomas?" Children responded verbally or by pointing.

same name and shape. The two objects differed on the critical feature. In *adjective noun* trials, the critical feature was mentioned (e.g., one object was labeled “It’s a blicket” and the other was labeled “It’s a striped blicket”); in *noun* trials, the critical feature was not mentioned (e.g., one object was labeled “It’s a blicket” and the other was also labeled “It’s a blicket”).

After each trial, children were asked to make a judgment about the prevalence of the critical feature in the novel category. For instance, they were asked, “Let’s think about all of the blickets on this planet. How many blickets do you think are striped?” There were four options on the screen, each a cloud of six of the same shape of novel object, with differing proportions having the critical feature and in color (and the remaining objects without the feature and in grey). The options were Few (1/6 with feature), Some (2/6 with feature), Most (4/6 with feature), and Almost All (5/6 with feature). After asking the question, the experimenter said the options: “Few of the blickets, some of the blickets, most of the blickets, or almost all of the blickets?” Children responded verbally. If they paused or seemed uncertain, the experimenter repeated the options. If the child preferred to point to the option on the screen (as happened with one participant), the experimenter asked the child’s parent to report the option they pointed to.

There were six trials in total. Half of trials were *adjective noun* trials and half were *noun* trials, and this factor was crossed with the feature type: size (wide or tall), color (blue or red), and pattern (spotted or striped). At each trial, the novel object shape and novel object name were randomly assigned out of a set of six names or shapes. The ordering of two objects in each trial (one with the critical feature and one without) was random.

Before the main task, children did two practice trials with familiar objects to establish that they understood the response measure. The two practice questions were: “Let’s think about all of the cookies in the world. How many cookies do you think are square?” and “Let’s think about all of the bananas in the world. How many bananas do you think are yellow?" They responded on the same scale used in the main task trials.

Performance on practice trials

Children's performance on the two practice trials with familiar objects can help give us a sense of whether they understand the typicality measure in this task. If the children understand this measure, we expect them to report that bananas are more commonly yellow than cookies are square. Out of 29 participants, 15 (0.52%) rated bananas to be more commonly yellow than cookies are square (0.40% of 5-year-olds and 0.64% of 6-year-olds). That is, many children, especially the younger 5-year-olds, either did not understand this measure well or did not believe that cookies are not typically square and bananas are typically yellow. Below, we will report the results of the main task both for all children and, separately, for just the children who performed correctly on the familiar practice trials to see whether there is evidence for contrastive inference among children who understood the measure.

Results

Our key question is whether children make different inferences when an object's feature is mentioned than when it is not. To test this question, we fit a linear regression with children's prevalence choices as the outcome (coded as *few* = 1, *some* = 2, *most* = 3, and *almost all* = 4) and utterance type (*noun* vs. *adjective noun*), feature type (color, size, or pattern), and their interaction as predictors, as well as a random intercept by subject. The effect of utterance type was marginally significant: children's prevalence judgments were marginally lower when there was an adjective in the utterance (β = -0.55, t = -1.95, p = 0.05). Effects of feature type were not significant ($\beta_{pattern}$ = -0.45, t = -1.58, p = 0.12; β_{size} = -0.28, t = -0.98, p = 0.33), nor were interactions between utterance type and feature type ($\beta_{adjective-noun*pattern}$ = 0.66, t = 1.64, p = 0.10; $\beta_{adjective-noun*size}$ = 0.48, t = 1.21, p = 0.23). Overall, we find weak evidence that children infer that mentioned features are less typical. Children's prevalence judgments are shown in Figure 10.

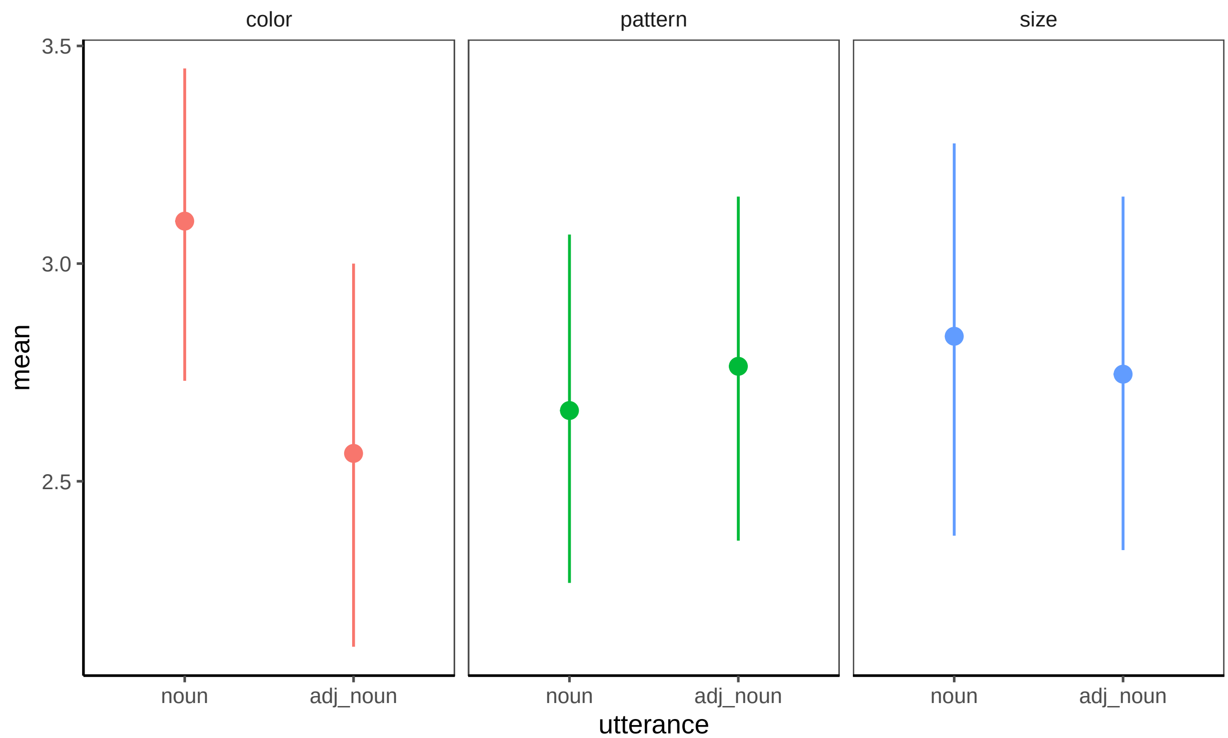

Figure 10. Children's prevalence judgments. Children directionally rate features to be less prevalent when they are mentioned with an adjective.

Based on their performance in the practice trials, it seems that many children did not understand the prevalence measure well. We can separately test the performance of children who correctly answered the practice trials to see whether children who understand the measure demonstrate contrastive inference. We fit the same model specification to only children who rated bananas to be yellow more typically than cookies are square. Among these children, there is a significant effect of utterance type, such that they infer that mentioned features are less typical (β = -1.13, t = -3.07, p = 0.00). Effects of feature type were not significant ($\beta_{pattern}$ = -0.47, t = -1.26, p = 0.21; β_{size} = -0.20, t = -0.54, p = 0.59), nor were interactions between utterance type and feature type ($\beta_{adjective-noun*pattern}$ = 0.87, t = 1.66, p = 0.10; $\beta_{adjective-noun*size}$ = 0.93, t = 1.79, p = 0.08). Children who performed correctly on familiar trials judged mentioned features to be less typical (Figure 11.

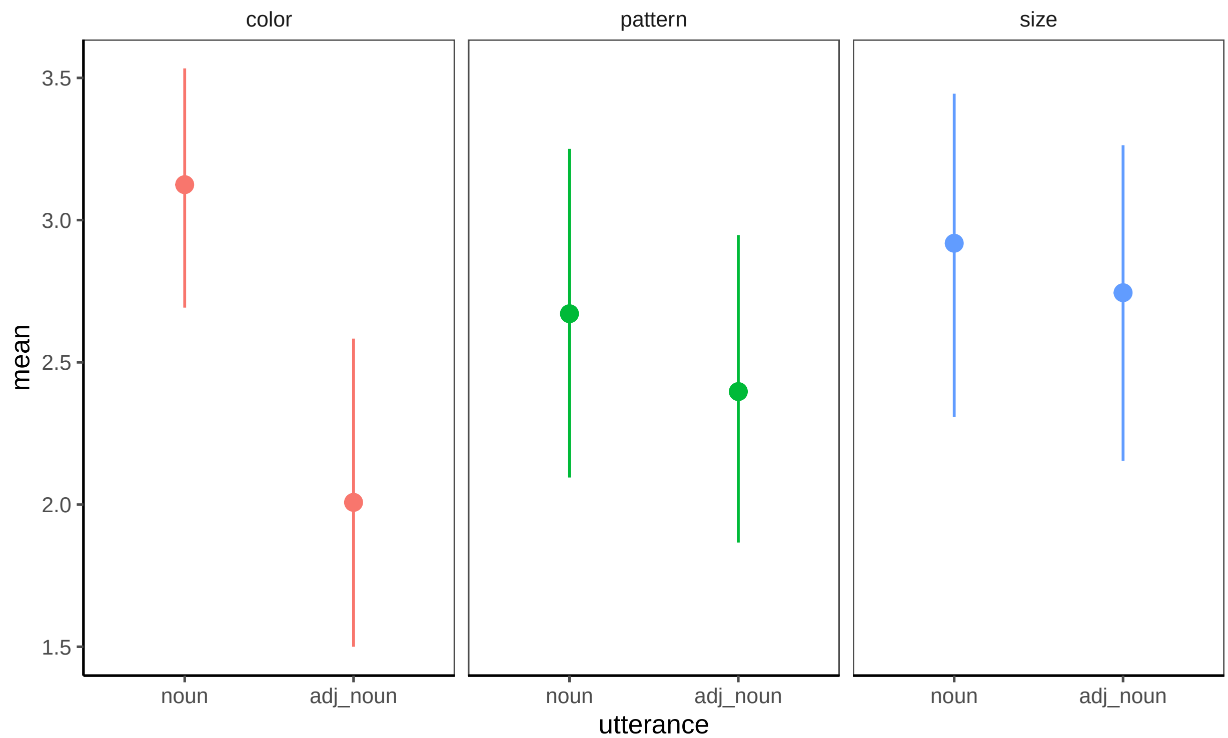

Figure 11. Prevalence judgments among only children who answered the practice trials correctly. These children rate features to be less prevalent when they are mentioned with an adjective.

Discussion

In this chapter, we ask how children develop the inference that that when a feature of a novel category is mentioned, that feature is likely to be atypical of the category. One possibility is that children simply associate the words, features and categories that are salient in an instance of reference. This would lead children to think a mentioned feature is representative or typical of the mentioned category. Another possibility is that children make the kind of contrastive inference adults make, inferring that the mentioned feature is atypical. In an exploratory study, we found suggestive evidence that 5–6-year-old children are not making an associative inference, and are directionally making a contrastive inference about mentioned features. Further, children who performed correctly on practice trials with familiar objects made significantly lower typicality judgments about mentioned features.

However, judging typicality is difficult for young children, and participants struggled with our measure overall. Evidence from this task is only preliminary, and calls for confirmatory tests with larger sample sizes and for the development of measures that are more sensible for young children.

Conclusion

This proposal maps out a course of research to understand how speakers selectively describe remarkable features and how listeners use this selective description to learn more about the world. In doing so, it inverts the framework that has positioned pragmatic inference as augmenting literal meaning that is already known, instead considering how people can use pragmatics to learn more about the semantics of unfamiliar things.

To understand how people use description to learn about the world, we first must know how description is used. Chapter 1 illustrates how caregivers use description in speaking to children, and proposes to more thoroughly test this question among adults speaking to adults. We find that parents predominantly mention atypical rather than typical features when speaking to children, and find the same pattern in our preliminary analysis of adult-adult speech. This pattern of description is consistent with the idea that people use language informatively with relation to background knowledge of the world, rather than giving veridical running commentary on the world's features. This finding raises questions about how children use description to learn, given that so many accounts of language learning rest on children forming associations among co-occurring words, features, and concepts. We also investigated whether language models that use associative learning among words can extract typical feature information from language. We find that these models do poorly in distinguishing between the typical and atypical features of nouns, with implications both for associative accounts of children's language learning and for language modeling. Overall, our findings highlight the complexity of learning about the world from language that describes it selectively.

However, perhaps people—unlike language models—know that language is used to selectively remark on the world, and can use this fact to learn about the unfamiliar. In Chapter 2, we investigated how adults make inferences about novel object categories, and found that they can use description to infer that a described feature is atypical. Further, even when description may have been used for another purpose—to establish reference—people still make inferences about typicality. We find that a model that considers the utility of utterances with respect to reference and typicality captures people's inferences. Much prior work has only considered the use of description in distinguishing between present referents (Engelhardt et al., 2011; Mangold & Pobel, 1988; Pechmann, 1989), and even work that has incorporated typicality has focused on reference as the primary goal of description (Mitchell et al., 2013; Rubio-Fernández, 2016a; Sedivy, 2003; Westerbeek et al., 2015a). Our findings emphasize that conveying typicality is likely a central factor in referring, and inferences about typicality are not secondary to or blocked by the purpose of establishing reference.

The ability to exploit description to learn more about the world than one has observed directly is most useful to people who are still rapidly learning—children. In Chapter 3, we propose to investigate how children make contrastive inferences about typicality, both of novel objects and novel social groups. The results of this experiment can tell us what children learn from the description they hear, as characterized in Chapter 1. Further, examining both novel objects and social groups can tell us how these inferences affect beliefs with social consequences, and whether children weigh direct observation against information from language differently in different domains.

The core computation in pragmatic inference is reasoning about alternatives—things the speaker could have said and did not. Given that others are reasoning about these alternatives, no choice is neutral. In the studies in Chapter 2, for instance, using an adjective in referring to an object led people to infer that the feature described by that adjective was less typical than if it had not been mentioned. But, conversely, *not* using an adjective led them to think that the feature was more typical than if they could not understand the meaning of the utterance at all—all communicative choices leak one's beliefs about the world. This has implications not only for learning about novel concrete objects, as people did here, but for learning about less directly accessible entities such as abstract concepts and social groups, as we propose to examine in Chapter 3. These inferences can be framed positively, as ways for learners to extract additional knowledge that was not directly conveyed, but can also spread beliefs that the speaker does not intend. The principle that people speak informatively is simple, but it holds unintuitive consequences—among speakers and listeners, humans and machines, adults and children—for describing and learning about the world.

Acknowledgements

Each chapter in this proposal represents collaborative work. The collaborators on each chapter are as follows: Chapter 1, Benjamin Morris and Dan Yurovsky; Chapter 2, Dan Yurovsky; Chapter 3, Rachel King and Dan Yurovsky. Thank you to my committee, Susan Goldin-Meadow, Marisa Casillas, Howard Nusbaum, and Dan Yurovsky for guidance throughout the preparation of this work.

Thank you to Ming Xiang, Benjamin Morris, Ashley Leung, Michael C. Frank, Judith Degen, Stephan Meylan, and Ruthe Foushee for feedback on portions of this research and manuscript. Portions of this work were published in the proceedings of Experiments in Linguistic Meaning (2020) and the proceedings of the 42nd annual meeting of the Cognitive Science Society. I am grateful for feedback from reviewers and attendees of Experiments in Linguistic Meaning, the meeting of the Cognitive Science Society, the meeting of the Society for Research in Child Development, the Midwestern Cognitive Science Conference, and the Dubrovnik Conference on Cognitive Science.

This research was funded by James S. McDonnell Foundation Scholar Award in Understanding Human Cognition #220020506 to Dan Yurovsky. The funding body had no involvement in the conceptualization, data collection, or analysis of this project.

References

Akhtar, N., Carpenter, M., & Tomasello, M. (1996). The Role of Discourse Novelty in Early Word Learning. *Child Development*, *67*(2), 635–645. https://doi.org/10.1111/j.1467-8624.1996.tb01756.x

Aparicio, H., Xiang, M., & Kennedy, C. (2016). Processing gradable adjectives in context: A visual world study. In *Semantics and linguistic theory* (Vol. 25, pp. 413–432).

Arts, A., Maes, A., Noordman, L. G. M., & Jansen, C. (2011). Overspecification in written instruction. *Linguistics*, *49*(3), 555–574.

Baillargeon, R. (1994). How do infants learn about the physical world? *Current Directions in Psychological Science*, *3*(5), 133–140.

Baker, N. D., & Greenfield, P. M. (1988). The development of new and old information in young children's early language. *Language Sciences*, *10*(1), 3–34.

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: The role of scalar alternatives in children's pragmatic inference. *Cognition*, *118*(1), 84–93.

Bedny, M., Koster-Hale, J., Elli, G., Yazzolino, L., & Saxe, R. (2019). There's more to "sparkle" than meets the eye: Knowledge of vision and light verbs among congenitally blind and sighted individuals. *Cognition*, *189*, 105–115.

Bergey, C., Morris, B., & Yurovsky, D. (2020). *Children hear more about what is atypical than what is typical*. PsyArXiv. https://doi.org/10.31234/osf.io/5wvu8

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, *51*(3), 437–457.

Boyer, T. W., Levine, S. C., & Huttenlocher, J. (2008). Development of proportional reasoning: Where young children go wrong. *Developmental Psychology*, *44*, 1478–1490. https://doi.org/10.1037/a0013110

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Clark, E. V. (1990). On the pragmatics of contrast. *Journal of Child Language*, *17*(2), 417–431. https://doi.org/10.1017/S0305000900013842

Dale, R., Kehoe, C., & Spivey, M. J. (2007). Graded motor responses in the time course of categorizing atypical exemplars. *Memory & Cognition*, *35*(1), 15–28.

Davies, C., Lingwood, J., Ivanova, B., & Arunachalam, S. (2021). Three-year-olds' comprehension of contrastive and descriptive adjectives: Evidence for contrastive inference. *Cognition*, *212*, 104707. https://doi.org/10.1016/j.cognition.2021.104707

Davies, M. (2008). The corpus of contemporary american english (coca): 520 million words, 1990-present.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, *127*, 591–621.

Degen, J., & Tanenhaus, M. K. (2015). Processing Scalar Implicature: A Constraint-Based Approach. *Cognitive Science*, *39*(4), 667–710.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint arXiv:1810.04805*.

Diesendruck, G., Hall, D. G., & Graham, S. A. (2006). Children's Use of Syntactic and Pragmatic Knowledge in the Interpretation of Novel Adjectives. *Child Development*, *77*(1), 16–30.

Engelhardt, P. E., Barış Demiral, Ş., & Ferreira, F. (2011). Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, *77*(2), 304–314. https://doi.org/10.1016/j.bandc.2011.07.004

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*.

Foppolo, F., Guasti, M. T., & Chierchia, G. (2012). Scalar Implicatures in Child Language: Give Children a Chance. *Language Learning and Development*, *8*(4), 365–394.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Frank, M. C., & Goodman, N. D. (2014). Inferring word meanings by assuming that speakers are informative. *Cognitive Psychology*, *75*, 80–96.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*(5), 578–585.

Gelman, S. A., & Markman, E. M. (1985). Implicit contrast in adjectives vs. Nouns: Implications for word-learning in preschoolers*. *Journal of Child Language*, *12*(1), 125–143.

Goldin-Meadow, S., Levine, S. C., Hedges, L. V., Huttenlocher, J., Raudenbush, S. W., & Small, S. L. (2014). New evidence about language and cognitive development based on a longitudinal study: Hypotheses for intervention. *American Psychologist*, *69*(6), 588.

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

Harris, P. L., & Koenig, M. A. (2006). Trust in testimony: How children learn about science and religion. *Child Development*, *77*(3), 505–524.

Horowitz, A. C., & Frank, M. C. (2016). Children's Pragmatic Inferences as a Route for Learning About the World. *Child Development*, *87*(3), 807–819.

Huang, Y. T., & Snedeker, J. (2008). Use of referential context in children's language processing. *Proceedings of the 30th Annual Meeting of the Cognitive Science Society.*

Huang, Y. T., & Snedeker, J. (2009). Semantic meaning and pragmatic interpretation in 5-year-olds: Evidence from real-time spoken language comprehension. *Developmental Psychology*, *45*(6), 1723–1739.

Johns, B. T., & Jones, M. N. (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, *4*(1), 103–120.

Kanwisher, N., Woods, R. P., Iacoboni, M., & Mazziotta, J. C. (1997). A locus in human extrastriate cortex for visual shape analysis. *Journal of Cognitive Neuroscience*, *9*(1), 133–142.

Landau, B., Gleitman, L. R., & Landau, B. (2009). *Language and experience: Evidence from the blind child* (Vol. 8). Harvard University Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

Legare, C. H., & Harris, P. L. (2016). The ontogeny of cultural learning. *Child Development*, *87*(3), 633–642.

Lewis, M., Zettersten, M., & Lupyan, G. (2019). Distributional semantics as a source of visual knowledge. *Proceedings of the National Academy of Sciences*, *116*(39), 19237–19238.

Mangold, R., & Pobel, R. (1988). Informativeness and Instrumentality in Referential Communication. *Journal of Language and Social Psychology*, *7*(3-4), 181–191.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the international conference on language resources and evaluation (lrec 2018)*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mitchell, M., Reiter, E., & Deemter, K. van. (2013). Typicality and Object Reference, 7.

Ni, W. (1996). Sidestepping garden paths: Assessing the contributions of syntax, semantics and plausibility in resolving ambiguities. *Language and Cognitive Processes*, *11*(3), 283–334.

Noveck, I. A. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: Experiments at the semantics–pragmatics interface. *Cognition*, *86*(3), 253–282.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics*, *27*(1), 89–110.

Rhodes, M., Leslie, S.-J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, *109*(34), 13526–13531.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 491.

Rubio-Fernández, P. (2016a). How Redundant Are Redundant Color Adjectives? An Efficiency-Based Analysis of Color Overspecification. *Frontiers in Psychology*, *7*.

Rubio-Fernández, P. (2016b). How Redundant Are Redundant Color Adjectives? An Efficiency-Based Analysis of Color Overspecification. *Frontiers in Psychology*, *7*.

Ryskin, R., Kurumada, C., & Brown-Schmidt, S. (2019). Information integration in modulation of pragmatic inferences during online language comprehension. *Cognitive Science*, *43*(8), e12769.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 45–50). Valletta, Malta: ELRA.

Sedivy, J. C. (2003). Pragmatic Versus Form-Based Accounts of Referential Contrast: Evidence for Effects of Informativity Expectations. *Journal of Psycholinguistic Research*, *32*(1), 3–23.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147.

Sloutsky, V. M., & Fisher, A. V. (2004). Induction and categorization in young children: A similarity-based model. *Journal of Experimental Psychology: General*, *133*(2), 166.

Snow, C. E. (1972). Mothers' speech to children learning language. *Child Development*, 549–565.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142).

Citeseer.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

Westerbeek, H., Koolen, R., & Maes, A. (2015a). Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00935

Westerbeek, H., Koolen, R., & Maes, A. (2015b). Stored object knowledge and the production of referring expressions: The case of color typicality. *Frontiers in Psychology*, *6*.

Willits, J. A., Sussman, R. S., & Amato, M. S. (2008). Event knowledge vs. Verb knowledge. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 2227–2232).

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological Review*, *114*(2), 245.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414–420.