Children hear more about what is atypical than what is typical

Claire Augusta Bergey*[1], Ben Morris*[2], & Dan Yurovsky[3]

[1] Stanford University

[2] Yale University

[3] Carnegie Mellon University

Author Note

Correspondence concerning this article should be addressed to Claire Augusta
Bergey*, 450 Jane Stanford Way, Building 420, Stanford, CA 94305. E-mail:
cbergey@stanford.edu

10   Abstract

11   How do children learn the typical features of things in the world? For many objects, this

12   information must come from the language they hear. However, language often does not

13   veridically reflect the world. If the language children hear selectively picks out the atypical

14   features of things (e.g., "purple carrot") rather than the typical features of things (e.g.,

15   "orange carrot"), learning about the world from language is less straightforward. Here, we

16   examine whether the language children hear from parents, as well as everyday conversation

17   among adults, overrepresents atypical features. We examined the typicality of features in

18   adjective-noun pairs produced by parents in a large, longitudinal corpus of parent-child

19   interaction, as well as a comparison set of adjective-noun pairs from adult-adult speech.

20   Across over 6,000 unique adjective–noun pairs, we found that parents speaking to

21   children—like adults speaking to other adults—predominantly use adjectives to mark

22   atypical features of things. We also found that parents of very young children comment on

23   typical features slightly more often than parents of older children. Language is structured

24   to emphasize what is atypical—so how can one learn about what things are typically like

25   from language? We also show that distributional semantics models that use word

26   co-occurrence to derive word meaning (word2vec) do not capture the typicality of

27   adjective–noun pairs well. A much more sophisticated language model (GPT-3) does

28   capture the typicality of adjective noun pairs well; though this model has input unlike what

29   children have access to, it provides useful bounds on the typicality information learnable

30   from applying simple training objectives to language alone. Taken together, our study

31   raises new foundational questions about how children manage to learn so much from

32   language that does not directly reflect the world, but selectively picks out remarkable

33   facets of it.

34   *Keywords:* language input, language acquisition, child-directed speech, corpus

35   analysis, language models

³⁶ Children hear more about what is atypical than what is typical

³⁷ Does language reflect the world? A strong correspondence between the world and
³⁸ language undergirds current theories of language and concept learning across a variety of
³⁹ domains. Children's early word learning is thought to proceed largely through dependable
⁴⁰ associations between language and sensory percepts (e.g., hearing "cup" and seeing a cup
⁴¹ at the same time) and words with other conceptually related words (e.g., associating "cup"
⁴² and "bowl" after hearing them together in an utterance) (Savic, Unger, & Sloutsky, 2022b;
⁴³ Sloutsky & Fisher, 2004; Smith & Yu, 2008; Unger, Savic, & Sloutsky, 2020a; Woodward,
⁴⁴ Markman, & Fitzsimmons, 1994). Congenitally blind children and adults learn visual
⁴⁵ concepts that are similar to those of their sighted peers, presumably primarily through
⁴⁶ language (Bedny, Koster-Hale, Elli, Yazzolino, & Saxe, 2019; Kim, Elli, & Bedny, 2019;
⁴⁷ Landau, Gleitman, & Landau, 2009). Further, language models' broad success in
⁴⁸ approximating human judgments across a variety of domains suggests that language
⁴⁹ supplies a lot of information about the world (Brown et al., 2020; Devlin, Chang, Lee, &
⁵⁰ Toutanova, 2018; Landauer & Dumais, 1997; Mikolov, Sutskever, Chen, Corrado, & Dean,
⁵¹ 2013).

⁵² In this paper, we argue that language in fact systematically departs from reflecting
⁵³ the world by selectively picking out remarkable facets of it. We rarely use language to
⁵⁴ provide running commentary on the world around us; instead, we use language to talk
⁵⁵ about things that diverge from our expectations or those of our conversational partner
⁵⁶ (Clark, 1990; Grice, 1975; Rohde, Futrell, & Lucas, 2021; Sperber, 1986). For instance, in
⁵⁷ lab tasks, people often mention the color of a brown banana but let the color of a yellow
⁵⁸ banana go unmentioned (Rubio-Fernández, 2016; Westerbeek, Koolen, & Maes, 2015).
⁵⁹ Given the communicative pressure to be informative, naturalistic language statistics may
⁶⁰ provide surprisingly little evidence about what is typical: we may rarely hear that a banana
⁶¹ is yellow. Here, we show that this pressure pervasively structures naturalistic language

utterance	pair	rating 1	rating 2	rating 3	mean
especially with wooden shoes.	wooden-shoe	2	2	2	2.00
you like red onions?	red-onion	5	3	4	3.60
the garbage is dirty.	dirty-garbage	7	6	6	6.00

Table 1

Sample typicality ratings from three human coders for three adjective-noun pairs drawn from the corpus. Ratings are on a scale from 1 (never) to 7 (always). Note that means may be slightly different from the mean of the three ratings shown here because some pairs have more than three ratings.

use—among adults, from adults to children, and by children—and complicates the problem faced by children and computational models when learning about the world from language.

To investigate whether people tend to mention the atypical, we first examined the typicality of adjectives with respect to the nouns they describe in a large corpus of adults' naturalistic conversation. We show that people's tendency to mention atypical features, as observed in constrained lab tasks, pervasively structures language use in a corpus of adults' conversations: people more often mention the atypical than the typical features of things.

We next examine whether parents, too, talk predominantly about the atypical features of things. If parents speak to children the way they speak to other adults, children may be faced with input that emphasizes atypicality in relation to world knowledge they do not yet have. On the other hand, parents may speak to children far 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. In a large corpus of parent-child interactions recorded in children's homes, we find that parents overwhelmingly choose to mention atypical rather than typical features; further, we find that children themselves mention more atypical than typical features.

78   We then ask whether the co-occurrence structure of language nonetheless captures

79 typicality information by testing whether the distributional semantics model word2vec

80 captures adjective-noun typicality. We find that relatively little typical feature information

81 is represented in these semantic spaces. We also test whether two more advanced language

82 models, BERT and GPT-3, capture typicality, and find that only the latter does well.

83 These models are unlikely to reflect children's learning mechanisms or language input, but

84 tell us what kinds of typicality information are learnable from language in principle.

Part I: People remark on the atypical

Method

87   In order to determine whether people use adjectives mostly to mark atypical features

88 of categories, we analyzed speech from large corpora of everyday conversations: adult-adult

89 conversations, caregivers' speech to children, and children's own speech. We extracted

90 adjectives and the nouns they modified from conversational speech, and asked a sample of

91 Amazon Mechanical Turkers to judge how typical the property described by each adjective

92 was for the noun it modified. We then examined both the broad features of this typicality

93 distribution and the way it changes over development.ˆ{Our typicality elicitation method,

94 and analyses and predicted results regarding child-directed speech were pre-registered at

95 the following link: https://osf.io/ypdzv/?view_only=4d34d324f2964336a28d8fa4b43e5580 .

96 This pre-registration specifies a prior version of our method for extracting adjective-noun

97 pairs, and results from the exact pre-registered analyses are available in a proceedings

98 paper [redacted for blind review], and conform to our pre-registered predictions. The

99 analyses in the present manuscript use an improved method for extracting adjective-noun

100 pairs, and conform to those same predictions. We did not pre-register the extraction

101 method we use in the present work because the corpus, analysis plan, and predictions

102 about child-directed speech did not change from the first pre-registration. The fact that

103 the same hypotheses are borne out under both extraction methods demonstrates that these

104 findings are robust to these data processing decisions.}

105   **Corpora.**   For adult-adult speech, we used data from the Conversation Analytic

106 British National Corpus, a corpus of naturalistic, informal conversations in people's

107 everyday lives (Albert, Ruiter, & Ruiter, 2015; Coleman, Baghai-Ravary, Pybus, & Grau,

108 2012). We excluded any conversations with child participants, for a total of 99,305

109 adult-adult utterances.

110   For our child-directed and child-produced speech, we used data from the Language

111 Development Project, a large-scale, longitudinal corpus of parent-child interactions

112 recorded in children's homes. Families were recruited to be representative of the

113 Chicagoland area in both socio-economic and racial composition; all families spoke English

114 at home (Goldin-Meadow et al., 2014). Recordings were taken in the home every 4 months

115 from when the child was 14 months old until they were 58 months old, resulting in 12

116 timepoints. Each recording was of a 90-minute session in which parents and children were

117 free to behave and interact as they liked. Our sample consisted of 64 typically developing

118 children and their caregivers with data from at least 4 timepoints ($mean = 11.3$

119 timepoints). Together, this resulted in a total of 641,402 parent utterances and 368,348

120 child utterances.

121   **Stimulus Selection.**   We parsed each utterance in our corpora using UDPipe, an

122 automated dependency parser, and extracted adjectives and the nouns they modified. This

123 set contained a number of abstract or evaluative adjective-noun pairs whose typicality

124 would be difficult to classify (e.g., "good"–"job"; "little"–"bit"). To resolve this issue, we

125 used human judgments of words' concreteness to identify and exclude non-concrete

126 adjectives and nouns (Brysbaert, Warriner, & Kuperman, 2014). We retained for analysis

127 only pairs in which both the adjective and noun were in the top 25% of concreteness

128 ratings (e.g., "dirty" – "dish"; "green" – "fish"). Additionally, one common adjective that

129 is used abstractly and evaluatively in British English but is concrete in American English

130 (*bloody*) was excluded from the set of pairs from the CABNC.

131   Our final sample included 6,370 unique adjective-noun pairs drawn from 7,471 parent

132 utterances, 2,775 child utterances, and 1,867 adult-adult utterances. The pairs were

133 combinations of 1,498 distinct concrete nouns and 1,388 distinct concrete adjectives. We

134 compiled these pairs and collected human judgments on Amazon Mechanical Turk for each

135 pair, as described below. Table 1 contains example utterances from the final set and

136 typicality judgments from our human raters.

Participants

138   Each participant rated 35 adjective-noun pairs, and we aimed for each pair to be

139 rated five times, for a total of 910 rating tasks. Participants were allowed to rate more than

140 one set of pairs and were paid $0.80 per task. Distribution of pairs was balanced using a

141 MongoDB database that tracked how often sets of pairs had been rated. If a participant

142 allowed their task to expire with the task partially complete, we included those ratings and

143 re-recruited the task. Overall, participants completed 32,461 ratings. After exclusions

144 using an attention check that asked participants to simply choose a specific number on the

145 scale, we retained 32,293 judgments, with each adjective–noun pair retaining at least two

146 judgments.

Design and Procedure

148   To evaluate the typicality of the adjective–noun pairs that appeared in parents'

149 speech, we asked participants on Amazon Mechanical Turk to rate each pair. Participants

150 were presented with a question of the form "How common is it for a cow to be a brown

151 cow?" and asked to provide a rating on a seven-point scale: (1) never, (2) rarely, (3)

152 sometimes, (4) about half the time, (5) often, (6) almost always, (7) always. We also gave

153 participants the option to select "Doesn't make sense" if they could not understand what

154 the adjective-noun pair would mean. Pairs that were marked with "Doesn't make sense" by

155 two or more participants were excluded from the final set of pairs: 1,591 pairs were

156 excluded at this stage, for a final set of 4,779 rated adjective-noun pairs. Some of these

157 nonsense pairs likely resulted from imperfect automated part of speech tagging (e.g.,

158 till—dinner, wipe—face); others were unorthodox uses of description or difficult to imagine

159 out of context (e.g., back—mom, square—circle, teeth—show). Though there are many of

160 these nonsense exclusions, this criterion is conservative and likely errs on the side of

161 excluding atypical pairs rather than typical ones.

162   **Results.**   We combined the human typicality ratings with usage data from our

163 corpora to examine the extent to which parents, children, and adults speaking to other

164 adults use language to describe typical and atypical features. In our analyses, we

165 token-weighted these judgments, giving higher weight to pairs that occurred more

166 frequently in speech. However, results are qualitatively identical and all significant effects

167 remain significant when examined on a type level.

168   First, we examine whether adults speaking to other adults in naturalistic

169 conversation talk about atypical features more than typical ones. Examining

170 adjective-noun usage in the Conversation Analytic British National Corpus, we found that

171 adult-adult speech predominantly features atypical adjective-pairs (Figure 1). To confirm

172 this effect statistically, we centered the ratings (i.e. "about half" was coded as 0), and then

173 predicted the rating on each trial with a mixed effects model with only an intercept and a

174 random effect of noun (`typicality ~ 1 + (1|noun)`). The intercept was reliably

175 negative, indicating that adult-adult speech more often points out atypical than typical

176 features (β = -0.94, t = -31.36, $p < .001$).

177   Though adults highlight atypical features when talking to other adults, they may

178 speak differently when talking to children. If caregivers speak informatively to convey what

179 is atypical or surprising in relation to their own sophisticated world knowledge, we should

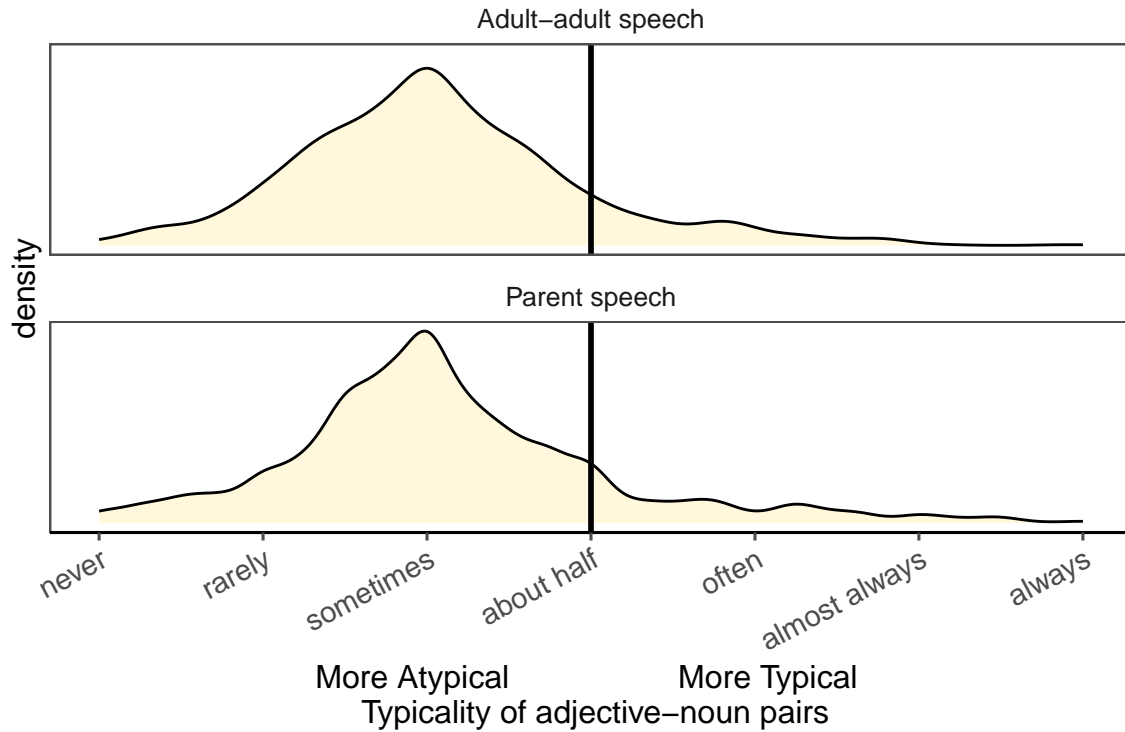180 see that caregiver description is dominated by adjectives that are sometimes or rarely true

Figure 1. Density plots showing use of atypical and typical adjective-noun pairs by parents speaking to children and adults speaking to other adults.

181 of the noun they modify. If instead child-directed speech privileges redundant information,

182 perhaps to align to young children's limited world knowledge, caregiver description should

183 yield a distinct distribution dominated by highly typical modifiers. Examining

184 adjective-noun use in the LDP, we found that caregivers' description predominantly focuses

185 on features that are atypical (Figure 2).

186   We confirmed this effect statistically using the same model structure as above, finding

187 a reliably negative intercept that indicates more atypical than typical adjective-noun pairs

188 (β = -0.85, t = -29.28, $p < .001$). We then re-estimated these models separately for each

189 age in the corpus, and found a reliably negative intercept for every age group (smallest

190 effect $\beta_{14-month-olds}$ = -0.69, t = -8.97, $p < .001$). Even when talking with very young

191 children, caregiver speech is structured according to the kind of communicative pressures

192 observed in adult-adult conversation.
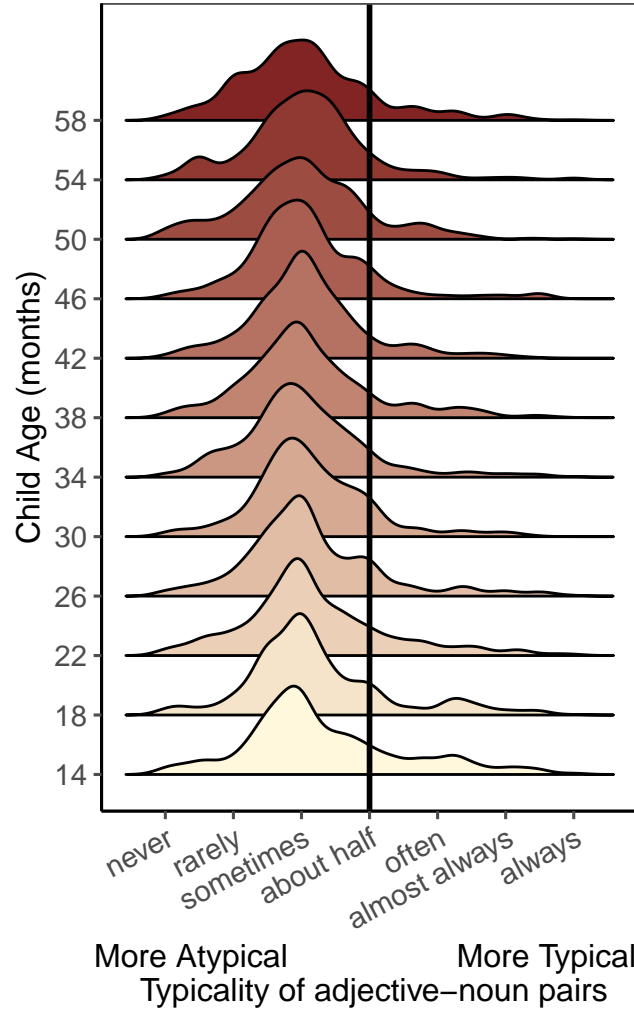
Figure 2. Density plots showing parents' use of atypical and typical adjective-noun pairs across their child's age.

₁₉₃   While description at every age tended to point out atypical features, this effect

₁₉₄ changed in strength over development. An age effect added to the previous model was

₁₉₅ reliably negative, indicating that parents of older children are relatively more likely to

₁₉₆ focus on atypical features (β = -0.09, t = -3.01, p = .003). In line with the idea that

₁₉₇ caregivers adapt their speech to their children's knowledge, it seems that caregivers are

₁₉₈ more likely to provide description of typical features for their young children, compared

₁₉₉ with older children. As a second test of this idea, we defined adjectives as highly typical if

₂₀₀ Turkers judged them to be 'often', 'almost always', or 'always' true. We predicted whether
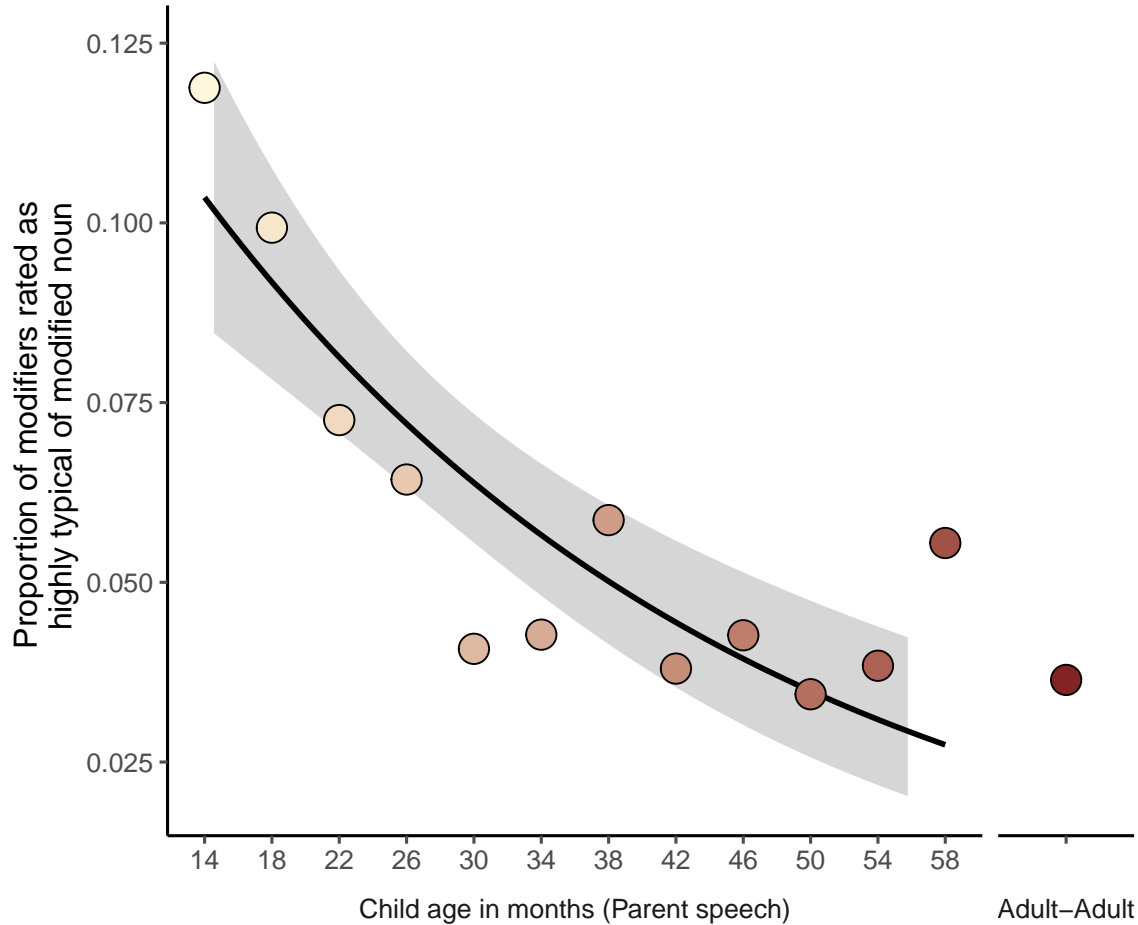
Figure 3. Proportion of caregiver description that is about highly typical features (often, almost always, or always true), as a function of age. Rightmost point: the proportion of description in adult-adult speech that is about highly typical features.

201 each judgment was highly typical from a mixed-effects logistic regression with a fixed effect

202 of age (log-scaled) and a random effect of noun. Age was a highly reliable predictor ($\beta =$

203 -0.69, $t =$ -3.80, $p < .001$). While children at all ages hear more talk about what is

204 atypically true (Figure 2), younger children hear relatively more talk about what is

205 typically true than older children do (Figure 3).

206   **Child Speech.**   Given the striking consistency in adult-to-adult speech and

207 caregiver speech across ages, we next consider what kind of information is contained in

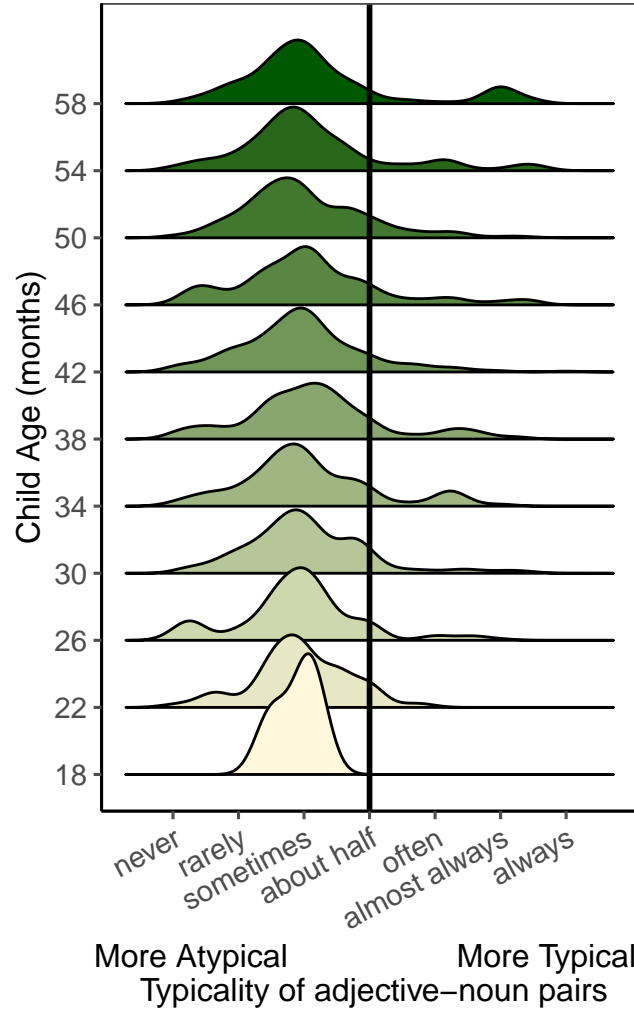208 children's speech. By analyzing children's own utterances, we can determine when children

Figure 4. Density plots showing children's use of atypical and typical adjective-noun pairs across age after excluding repeated utterances.

209 come to use description in a way that looks like adult speech. Are children mirroring

210 adult-like uses of description even from a young age, or are they choosing to describe more

211 typical features of the world?

212   We analyzed children's use of adjective–noun pairs and found that, following the

213 pattern of parent speech and adult-adult speech, they predominantly mention atypical

214 rather than typical features; confirmed statistically as above, we find a reliably negative

215 intercept (β = -0.96, t = -23.98, p < .001). One deflationary explanation for this pattern is

that children are simply often repeating the adjective-noun pairs their parents just

produced. To rule out this explanation, we re-analyzed the data excluding any

adjective-noun pairs produced by a parent in the past five utterances in conversation, still

finding a reliably negative intercept ($\beta =$ -0.97, $t =$ -22.31, $p <$ .001). Further, when

testing within each age group, even the 22-month-olds (the first age for which we have

sufficient child adjective-noun utterances to estimate) are reliably producing more atypical

than typical adjective-noun pairs; the intercept is reliably negative when estimated within

every age (14-month-olds and 18-month-olds are excluded due to having 0 and 3

adjective-noun pairs, respectively; estimate at 22 months old, $\beta =$ -1.07, $t =$ -8.36, $p <$

.001) That is, even when excluding utterances children may have immediately imitated

from their parents, and from the earliest ages they are consistently using adjective-noun

pairs, children more often mention atypical than typical features of things (Figure 4).

The fact that children are remarking on atypical features is intriguing, but it would

be premature to conclude that they are doing so to be selectively informative. Note also

that especially at young ages, children produce few adjective-noun pairs—they are not

producing any at 14 months old, our earliest timepoint—so our data on children's speech is

somewhat sparse. We discuss potential interpretations of this finding further in the

Conclusion.

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

242 describes atypical features.

243   It should be noted that children's utterances come from naturalistic conversations

244 with caregivers, and their use of atypical description may be prompted by parent-led

245 discourse. That is, if a caregiver chooses to describe the *purpleness* of a cat in book, the

246 child may well respond by asking about that same feature. Further, atypical descriptors

247 may actually be more likely to elicit imitation from child speakers, compared with typical

248 descriptors (Bannard, Rosner, & Matthews, 2017). Future analyses would need to better

249 disentangle the extent to which children's productions are imitative of caregivers.

250   This usage pattern aligns with the idea that language is used informatively in relation

251 to background knowledge about the world. It may pose a problem, however, for young

252 language learners with still-developing world knowledge. If language does not transparently

253 convey the typical features of objects, and instead (perhaps misleadingly) notes the

254 atypical ones, how might children come to learn what objects are typically like? One

255 possibility is that information about typical features is captured in more complex

256 regularities across many utterances. If this is true, language may still be an important

257 source of information about typicality as children may be able to extract more accurate

258 typicality information by tracking statistical regularities across many utterances.

Extracting Typicality from Language Structure

260   Much information can be gleaned from language that does not seem available at first

261 glance. From language alone, simple distributional learning models can recover enough

262 information to perform comparably to non-native college applicants on the Test of English

263 as a Foreign Language (Landauer & Dumais, 1997). Recently, Lewis, Zettersten, and

264 Lupyan (2019) demonstrated that even nuanced feature information may be learnable

265 through distributional semantics alone, without any complex inferential machinery.

266 Further, experiments with adults and children suggest that co-occurrence regularities may

267 help structure semantic knowledge (Savic, Unger, & Sloutsky, 2022a, 2023; Unger, Savic, &

268 Sloutsky, 2020b). Relationships among nouns that reflect feature information such as size

269 are recoverable in the semantic spaces of large language models (Grand, Blank, Pereira, &

270 Fedorenko, 2022). However, language models show deficits in inferring typicality and

271 atypicality in more controlled tasks, departing systematically from human-like pragmatic

272 inference (Kurch, Ryzhova, & Demberg, 2024; Misra, Ettinger, & Rayz, 2021).

273   Here, we ask whether a simple distributional semantics model trained on the

274 language children hear can capture typical feature information. Further, we test whether a

275 distributional semantics model trained on a larger corpus of adult-directed text as well as

276 two more sophisticated language models capture adjective-noun typicality. These models

277 are trained on more and different language than is available to children, but tell us more

278 about whether and how typicality information is learnable by applying simple learning

279 objectives to text.

Method

281   To test the possibility that simple distributional semantics models would capture

282 typicality relationships between nouns and adjectives, we trained word2vec on the same

283 corpus of child-directed speech used in our first set of analyses. Word2vec is a neural

284 network model that learns to predict words from the contexts in which they appear. This

285 leads word2vec to encode words that appear in similar contexts as similar to one another

286 (Firth, 1957).

287   We used the continuous-bag-of-words (CBOW) implementation of word2vec in the

288 `gensim` package (Řehůřek & Sojka, 2010). We trained the model using a surrounding

289 context of 5 words on either side of the target word and 100 dimensions (weights in the

290 hidden layer) to represent each word. After training, we extracted the hidden layer

291 representation of each word in the model's vocabulary—these are the vectors used to

292 represent these words.

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

300   While word2vec straightforwardly represents what can be learned about word
301 similarity by associating words with similar contexts, it does not represent the cutting edge
302 of language modeling. Perhaps more sophisticated models trained on larger corpora would
303 represent these typicalities better. To test this, we asked how BERT (Devlin et al., 2018)
304 and GPT-3 (Brown et al., 2020) represent typicality. BERT is a masked language model
305 trained on BookCorpus and English Wikipedia, which represents the probability of words
306 occurring in slots in a phrase. We gave BERT phrases of the form "_____ apple", and
307 asked it the probability of different adjectives filling the empty slot.

308   GPT-3 is a generative language model trained on large quantities of internet text,
309 including Wikipedia, book corpora, and web page text from crawling the internet. Because
310 it is a generative language model, we can ask GPT-3 the same question we asked human
311 participants directly and it can generate a text response. We prompted the
312 `davinci-text-003` instance of GPT-3 questions of the form: "You are doing a task in
313 which you rate how common it is for certain things to have certain features. You respond
314 out of the following options: Never, Rarely, Sometimes, About half the time, Often, Almost
315 always, or Always. How common is it for a cow to be a brown cow?" Because BERT and
316 GPT-3 are trained on more and different kinds of language than what children hear, results
317 from these models likely do not straightforwardly represent the information available to

318 children in language. However, results from BERT and GPT-3 can indicate the challenges

319 language models face in representing world knowledge when the language people use

320 emphasizes remarkable rather than typical features.

321 **Results**

322   We find that similarities in the model trained on the Language Development Project

323 corpus have near zero correlation with human adjective–noun typicality ratings ($r = 0.05$,

324 $p = .001$). However, our model does capture other meaningful information about the

325 structure of language, such as similarity within part of speech categories. Comparing with

326 pre-existing large-scale human similarity judgements for word pairs, our model shows

327 significant correlations (correlation with wordsim353 similarities of noun pairs, 0.28;

328 correlation with simlex similarities of noun, adjective, and verb pairs, 0.16). This suggests

329 that statistical patterns in child-directed speech are likely insufficient to encode

330 information about the typical features of objects, despite encoding at least some

331 information about word meaning more broadly.

332   However, the corpus on which we trained this model was small; perhaps our model

333 did not get enough language to draw out the patterns that would reflect the typical

334 features of objects. To test this possibility, we asked whether word vectors trained on a

335 much larger corpus—English Wikipedia—correlate with typicality ratings. This model's

336 similarities were significantly correlated with human judgments, although the strength of

337 the correlation was still fairly weak ($r = 0.34$, $p < .001$). How do larger and more

338 sophisticated language models fare? Like Wikipedia-trained word2vec, BERT's

339 probabilities were significantly correlated with human judgments, though weakly so ($r = $

340 0.15, $p < .001$). However, GPT-3's ratings were much better aligned with human

341 judgments ($r = 0.57$, $p < .001$).

342   Similarity judgments produced by our models reflect many dimensions of similarity,

but our human judgments reflect only typicality. To account for this fact and control for semantic differences among the nouns in our set, we performed a second analysis in which we considered only the subset of 109 nouns that had both a high-typicality (rated as at least "often") and a low-typicality (rated as at most "sometimes") adjective. We then asked whether the word2vec models rated the high-typicality adjective as more similar to the noun it modified than the low-typicality adjective. The LDP model correctly classified 49 out of 109 (44.95%), which was not different from chance ($p = .338$). The Wikipedia-trained word2vec model correctly classified 84 out of 109 (77.06%), which was better than chance according to a binomial test, though not highly accurate ($p < .001$). Figure 5 shows the word2vec models' similarities for the 109 nouns and their typical and atypical adjectives alongside scaled average human ratings.

The analogous analysis on BERT asks whether the model rates the high-typicality adjective as more likely to come before the noun than the low typicality adjective (e.g., P("red") > P("brown") in "_____ apple"). BERT correctly classified 66 out of 109 (60.55%), which is significantly better than chance ($p = .035$). However, BERT's performance was directionally less accurate than Wikipedia-trained word2vec: though BERT is a more sophisticated model, it does not capture adjective-noun typicality better than word2vec in this analysis. GPT-3 performs much better than BERT and the word2vec models, with 96 out of 109 (88.07%; $p < .001$). Figure 6 shows BERT and GPT-3 ratings for the 109 nouns and their typical and atypical adjectives alongside scaled average human ratings.

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 allows people to
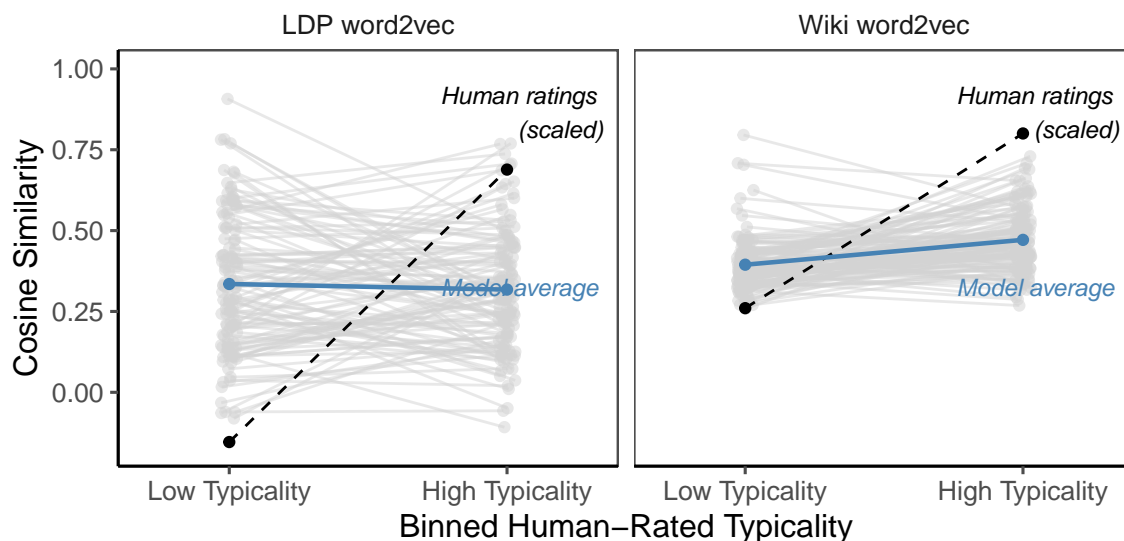
Figure 5. Plots of word2vec noun-adjective similarities for nouns for which there was at least one atypical adjective (rated at most "sometimes"), and at least one typical adjective (rated at least "often"). Human ratings line depicts the mean human rating in each group, scaled to the range of model outputs.

converse informatively 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, 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 two word2vec models, one trained on child-directed language and one trained on adult-adult language, did not capture typicality very well. Even BERT, a language model trained on much more text and with a more complex architecture, did not perform better than a Wikipedia-trained word2vec model in
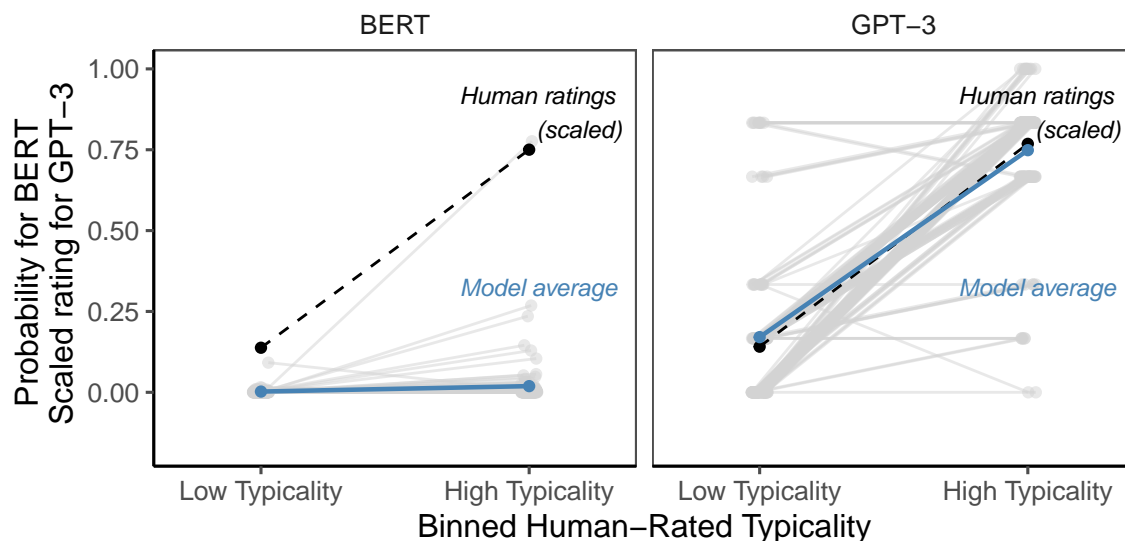
Figure 6. Plots of BERT and GPT-3 noun-adjective similarities for nouns for which there was at least one atypical adjective (rated at most "sometimes"), and at least one typical adjective (rated at least "often"). Human ratings line depicts the mean human rating in each group, scaled to the range of model outputs.

reflecting typicality. This may be because these models are designed to capture language statistics, with BERT in particular capturing which words are likely to occur following one another—and as we show in our corpus analyses, adjective-noun pairs that come together often reflect atypicality rather than typicality. Note that a consistent *inverse* relationship—rating high-typicality pairs as *less* similar or *less* probable—would also be evidence that these models capture typicality, but the word2vec models and BERT do not evince this pattern either. However, GPT-3 captured typicality quite well, suggesting that the way people structure language to emphasize atypicality is not necessarily an impediment for much larger models' representation of typicality. Further work remains to understand how GPT-3 comes to represent typicality relationships so much better than the smaller models we tested. Overall, a large language model trained on text much greater in quantity and different in quality from child-directed language did capture adjective-noun typicality well, but models with simpler learning mechanisms and language input more

395 similar to what is available to children did not.

396   Of course, perceptual information from the world may simplify the problem of

397 learning about typicality. In many cases, perceptual information may swamp information

398 from language; children likely see enough orange carrots in the world to outweigh hearing

399 "purple carrot." It remains unclear, however, how children learn about categories for which

400 they have scarcer evidence. Indeed, language information likely swamps perceptual

401 information for many other categories, such as abstract concepts or those that cannot be

402 learned about by direct experience. If such concepts pattern similarly to the concrete

403 objects analyzed here, children are in a particularly difficult bind.

404   It is also possible that other cues from language and interaction provide young

405 learners with clues to what is typical or atypical, and these cues are uncaptured by our

406 measure of usage statistics. Caregivers may highlight when a feature is typical by using

407 certain syntactic constructions, such as generics (e.g., "tomatoes are red"), and children

408 may learn especially well from rarer constructions that use adjectives postnominally or

409 contrast among referents present in the discourse context (Au & Markman, 1987; Davies,

410 Lingwood, & Arunachalam, 2020; Waxman & Klibanoff, 2000). Caregivers may also mark

411 the atypicality of a feature using extralinguistic cues, e.g., by demonstrating surprise using

412 prosody and facial expressions. Such cues from language and interaction may provide key

413 cues to interpretation; however, given the sheer frequency of atypical descriptors, it seems

414 unlikely that they are consistently well-marked.

415   Another possibility is that children expect language to be used informatively at a

416 young age. Under this hypothesis, their language environment is not misleading at all, even

417 without additional cues from caregivers. Children as young as two years old tend to use

418 words to comment on what is new rather than what is known or assumed (Baker &

419 Greenfield, 1988; Bohn, Tessler, Merrick, & Frank, 2021). Children may therefore expect

420 adjectives to comment on surprising features of objects. If young children expect adjectives

to mark atypical features (Horowitz & Frank, 2016), as adults do (Bergey & Yurovsky, 2023), they can use description and the lack thereof to learn more about the world. Our finding that children themselves mostly remark on atypical rather than typical features of things is consistent with this possibility, though does not provide strong evidence that children understand to use description informatively.

Whether adult-directed, child-directed, or a child's own speech, 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. The pressure for language to be informative is a pervasive force structuring the language children hear, and future work must disentangle whether children capitalize on or are misled by this selective informativity in learning about the world.

> Stimuli, data, and analysis code
> available at XXXXXXXX

References

Albert, S., Ruiter, L. E. de, & Ruiter, J. P. de. (2015). *CABNC: The Jeffersonian transcription of the Spoken British National Corpus.*

Au, T. K., & Markman, E. M. (1987). Acquiring word meanings via linguistic contrast. *Cognitive Development*, *2*(3), 217–236.

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.

Bannard, C., Rosner, M., & Matthews, D. (2017). What's worth talking about? Information theory reveals how children balance informativeness and ease of production. *Psychological Science*, *28*(7), 954–966.

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. A., & Yurovsky, D. (2023). Using contrastive inferences to learn about new words and categories. *Cognition*, *241*, 105597.

Bohn, M., Tessler, M. H., Merrick, M., & Frank, M. C. (2021). How young children integrate information sources to infer the meaning of words. *Nature Human Behaviour*, *5*(8), 1046–1054.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., . . . Amodei, D. (2020). *Language Models are Few-Shot Learners.* arXiv. https://doi.org/10.48550/arXiv.2005.14165

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.

Coleman, J., Baghai-Ravary, L., Pybus, J., & Grau, S. (2012). *Audio BNC: The audio edition of the Spoken British National Corpus.*

Davies, C., Lingwood, J., & Arunachalam, S. (2020). Adjective forms and functions in british english child-directed speech. *Journal of Child Language*, *47*(1), 159–185.

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

471 *arXiv:1810.04805.*

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

473 *Analysis.*

474 Goldin-Meadow, S., Levine, S. C., Hedges, L. V., Huttenlocher, J., Raudenbush, S. W., &

475 Small, S. L. (2014). New evidence about language and cognitive development based on

476 a longitudinal study: Hypotheses for intervention. *American Psychologist*, *69*(6), 588.

477 Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection

478 recovers rich human knowledge of multiple object features from word embeddings.

479 *Nature Human Behaviour*, *6*(7), 975–987.

480 Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

481 Horowitz, A. C., & Frank, M. C. (2016). Children's Pragmatic Inferences as a Route for

482 Learning About the World. *Child Development*, *87*(3), 807–819.

483 Kim, J. S., Elli, G. V., & Bedny, M. (2019). Knowledge of animal appearance among

484 sighted and blind adults. *Proceedings of the National Academy of Sciences*, *116*(23),

485 11213–11222. https://doi.org/10.1073/pnas.1900952116

486 Kurch, C., Ryzhova, M., & Demberg, V. (2024). Large language models fail to derive

487 atypicality inferences in a human-like manner. *Proceedings of the Workshop on*

488 *Cognitive Modeling and Computational Linguistics*, 86–100.

489 Landau, B., Gleitman, L. R., & Landau, B. (2009). *Language and experience: Evidence*

490 *from the blind child* (Vol. 8). Harvard University Press.

491 Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent

492 semantic analysis theory of acquisition, induction, and representation of knowledge.

493 *Psychological Review*, *104*(2), 211.

494 Lewis, M., Zettersten, M., & Lupyan, G. (2019). Distributional semantics as a source of

495 visual knowledge. *Proceedings of the National Academy of Sciences*, *116*(39),

496 19237–19238.

497 Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in

pre-training distributed word representations. *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. *Advances in Neural Information Processing Systems*, 3111–3119.

Misra, K., Ettinger, A., & Rayz, J. T. (2021). Do language models learn typicality judgments from text? *arXiv Preprint arXiv:2105.02987*.

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

Rohde, H., Futrell, R., & Lucas, C. G. (2021). What's new? A comprehension bias in favor of informativity. *Cognition*, *209*, 104491.

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

Savic, O., Unger, L., & Sloutsky, V. M. (2022a). Exposure to co-occurrence regularities in language drives semantic integration of new words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(7), 1064–1081. https://doi.org/10.1037/xlm0001122

Savic, O., Unger, L., & Sloutsky, V. M. (2022b). Exposure to co-occurrence regularities in language drives semantic integration of new words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(7), 1064.

Savic, O., Unger, L., & Sloutsky, V. M. (2023). Experience and maturation: The contribution of co-occurrence regularities in language to the development of semantic organization. *Child Development*, *94*(1), 142–158. https://doi.org/10.1111/cdev.13844

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.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via

525 cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

526 Sperber, D. (1986). *Relevance: Communication and cognition*. Blackwell.

527 Unger, L., Savic, O., & Sloutsky, V. M. (2020b). Statistical regularities shape semantic

528 organization throughout development. *Cognition*, *198*, 104190.

529 https://doi.org/10.1016/j.cognition.2020.104190

530 Unger, L., Savic, O., & Sloutsky, V. M. (2020a). Statistical regularities shape semantic

531 organization throughout development. *Cognition*, *198*, 104190.

532 Waxman, S. R., & Klibanoff, R. S. (2000). The role of comparison in the extension of

533 novel adjectives. *Developmental Psychology*, *36*(5), 571.

534 Westerbeek, H., Koolen, R., & Maes, A. (2015). Stored object knowledge and the

535 production of referring expressions: The case of color typicality. *Frontiers in*

536 *Psychology*, *6*.

537 Woodward, A. L., Markman, E. M., & Fitzsimmons, C. M. (1994). Rapid word learning

538 in 13-and 18-month-olds. *Developmental Psychology*, *30*(4), 553.