A computational theory of child overextension (supplementary material)

Renato Ferreira Pinto, Jr.[1], Yang Xu[2†]
[1]Department of Computer Science, University of Toronto
[2]Department of Computer Science, Cognitive Science Program, University of Toronto
[†]Correspondence: yangxu@cs.toronto.edu

Contents

List of Figures

List of Tables

1 Data and code repository

Data and code for replication including a demonstration are deposited at `https://github.com/r4ferrei/computational-theory-overextension`.

2 Sensitivity parameter and model simulation

We offer an illustration of the proposed overextension framework in a simulation of child naming behavior based on a range of values for the sensitivity parameter. We expect the modeling framework to reproduce qualitative characteristics of the development of children's productive speech, from overly broad extension patterns to more precise ways of word usage that approximate the conventional lexicon. In particular, we progressively shrink the kernel width or increase sensitivity in the parameter h, i.e., making the model more sensitive to the semantic appropriateness of word choices specified in the likelihood term of Equation 1 in the main text, which trades off with the cognitive effort specified in the frequency-based prior. Initially, we expect children to rely more on frequency prior in their strategies toward overextension. Over the course of the parameter variation, we expect children to rely more on the likelihood in overextension due to increasing sensitivity or precision in their semantic representation. We obtained the top word productions predicted by the model at different points in the simulated environment and recorded the rates of overextended vs conventional word usage over the range of parameter values we considered.

Figure S1a shows the predicted rate of overextension over the range of parameter values, measured by the proportion of words in the vocabulary predicted to overextend by the model. We observe a peak rate close to the parameter value estimated from empirical data, suggesting that our model captures the period of overextension reported in the developmental literature. To the left of the peak, there is a decrease in overextension rate, reflecting a more limited effective vocabulary as the model favours a core set of high-frequency words due to the pressure of cognitive effort. To the right of the peak, the opposite effect holds: as sensitivity increases in semantic space and becomes more dominant over the role of cognitive effort, overextension gives way to novel words and becomes increasingly rare, until it ceases to exist while the model predicts production to converge toward conventional lexical use.

Figure S1b shows concrete examples of top model productions in naming referents from three different domains, animals, fruits, and vehicles, over the course of simulation. We observe that, initially, words such as *car* and *apple* are broadly overextended and serve as central elements for their respective domains, echoing observations by Rescorla [3]. As the parameter shrinks in value, categories become progressively narrower even as some acquired words are overextended to similar referents (e.g., *orange* to lemons). Eventually, the productions converge toward conventional names, illustrating the ability of the model to capture precise ways of naming as well as overextension.

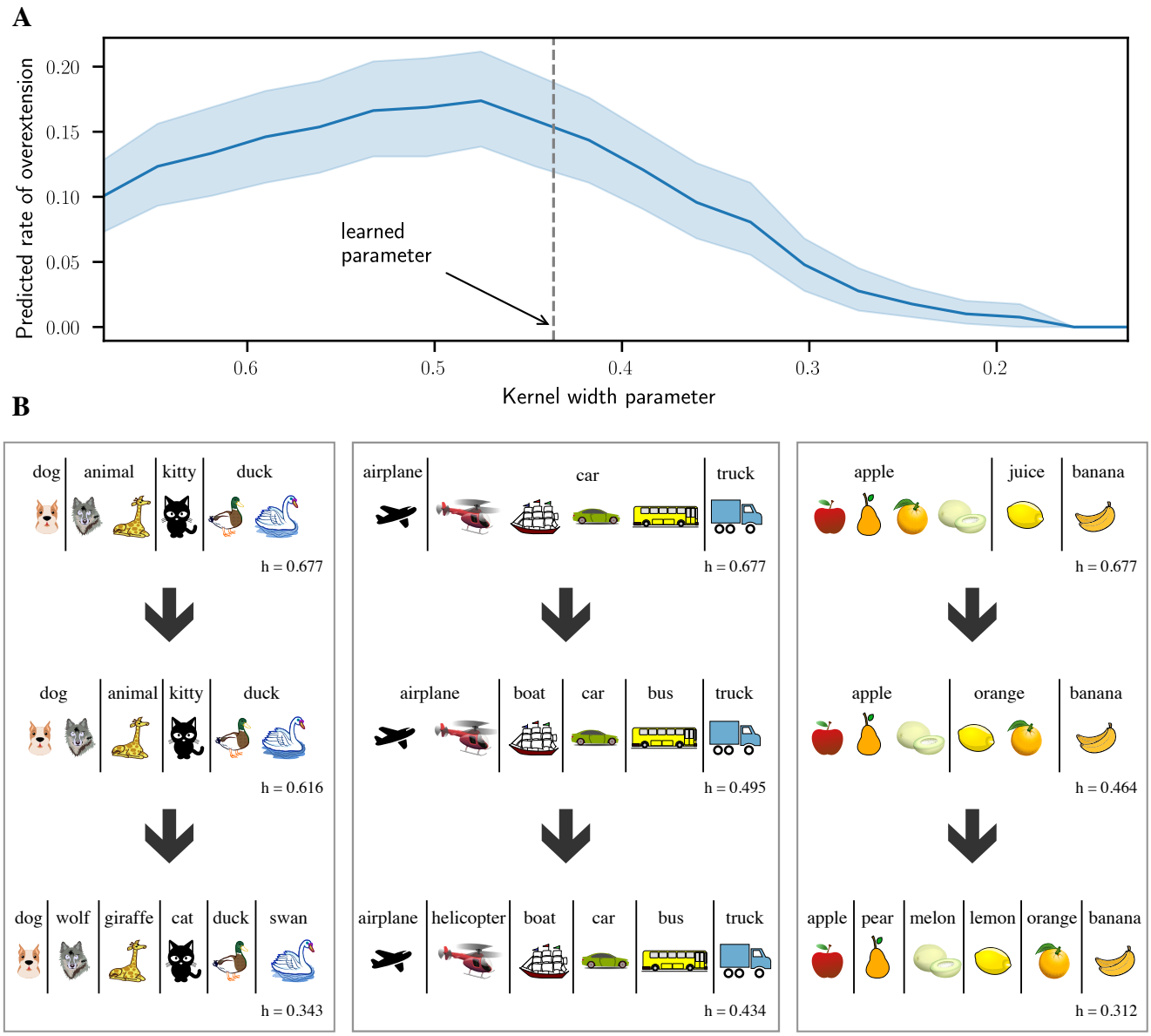

Figure S1: Results from model simulation over a range of values for the sensitivity parameter. a) Predicted rate of overextended words in the child vocabulary ($N = 317$). Shaded region represents bootstrap 95% confidence interval. b) Top model-produced child naming of sample concepts in three domains over the course of simulation. The top row reflects the trend under the largest parameter values, the middle row shows the intermediate range, and the bottom row shows the trend under the smallest parameter values. Each stage shows the kernel width (h) parameter value at which the model first produces that pattern.

3 Evidence for multimodality in overextension

To examine the amount of variability that the multimodal semantic space we constructed explains in the overextension data, we performed a logistic regression analysis that evaluates whether the model is able to discern overextended word-referent pairs with shuffled versions of these pairs. In particular, we considered two sets of data: the *attested set* of overextension word-referent pairs, and a *control set* that shuffles the word-referent mappings from the attested set. We then performed a binary classification task via logistic regression to assess whether the attested set could be distinguished from the control set, given the same three relational features used for the other analyses. For each word-referent pair, the logistic regression factors were the z-scores of categorical, visual analogical, and predicate-based distances, normalized over the entire dataset, and the response was a binary indicator for the attested/control set. We also labelled each word-referent pair in the attested set according to the top-scoring feature in the logistic regression model as a way to approximate which multimodal feature best explains each instance of overextension, and thus assess the degree of multimodality in the overextension data.

Table S1 shows the BIC scores and cross-validated accuracies of the full multimodal model, partial models consisting of feature pairs, and partial models consisting of single features. The full multimodal model best distinguishes attested overextension from control word pairs in BIC score and predictive accuracy. Furthermore, all three features of the full multimodal model reached significance in the logistic regression ($p < .001$, $N = 472$). These results suggest that a combination of semantic relations provides significant predictability of concept pairs that might undergo overextension.

Table S1: BIC scores and cross-validated accuracies of logistic regression ($N = 472$). Parenthesized items indicate standard errors.

Model	**BIC score**	**Accuracy**
categorical (cat.)	495	0.778(19)
visual (vis.)	464	0.767(19)
predicate (pred.)	469	0.763(20)
vis. + pred.	408	0.807(18)
cat. + vis.	393	0.816(18)
cat. + pred.	422	0.805(18)
all features	374	0.839(17)

4 Comparison of single and multi-parameter models

Section 6.1 in the main text compares models for predicting word choices in overextension using the full multimodal semantic space or under restrictions to single features or feature pairs, as well as when placing a frequency-based prior on words versus a uniform prior. The experimental results showed that both multimodal feature integration and frequency-based prior yielded superior predictive performance, confirming the hypotheses we set out to test.

Table S2: Bayesian Information Criterion (BIC) scores for production models with respect to overextension dataset ($N = 236$) under different choices of prior, semantic features, and model formulation (i.e., single versus multiple kernel width parameters). A lower BIC score indicates a better model.

Model	**BIC score**			
	frequency prior		**uniform prior**	
	single param.	**multiple param.**	**single param.**	**multiple param.**
baseline	2471	2471	2717	2717
categorical (cat.)	1863	1863	2093	2093
visual (vis.)	1817	1817	2041	2041
predicate (pred.)	1853	1853	2072	2072
vis. + pred.	1732	1690	1947	1904
cat. + vis.	1682	1648	1904	1869
cat. + pred.	1646	1650	1871	1874
all features	**1592**	**1583**	**1812**	**1799**

Given the parsimonious formulation of our models, a natural question is whether the same conclusions still apply under less constrained models. In this section, we replicate the experimental results for models that allow for multiple kernel width parameters—one per semantic feature—rather than a single parameter as in the main text. Concretely, we adapted the formulation from Section 3.2 by defining the concept similarity function as follows:

$$\text{sim}(c_1, c_2) = \exp\left(-\frac{d_c(c_1,c_2)^2}{h_c} - \frac{d_v(c_1,c_2)^2}{h_v} - \frac{d_p(c_1,c_2)^2}{h_p}\right) \tag{1}$$

Under this formulation, each psychological dimension is modulated by an independent parameter, and we jointly optimized the model to maximize the *a posteriori* probability of the word-referent overextension pairs as in Section 5.4 of the main text.

Table S2 shows the BIC scores of all models under each choice of formulation, word prior, and semantic features. The main conclusions from our analyses still hold: models incorporating semantic features performed better than the baseline (i.e., lower in BIC scores); models with the frequency-based prior outperformed those with uniform priors; and models with featural integration performed better than those with isolated features (i.e., all features $<$ feature pairs $<$ single features in BIC score). Furthermore, the BIC scores of multi-parametric models were only slightly lower than the corresponding single-parameter models, compared to the larger differences between models with different priors or combinations of semantic features (even across model formulations). This result further validates our choice to use the more parsimonious model in the rest of our analyses here and in the main text.

5 Detailed results for model prediction of word choices

5.1 Predictive performance of partial models

Section 6.1 in the main text shows the average performance curves of models containing single features, feature pairs, and all three multimodal features. Figure S2 shows the performance curves of all individual models, thus elucidating the relative performance of each combination.

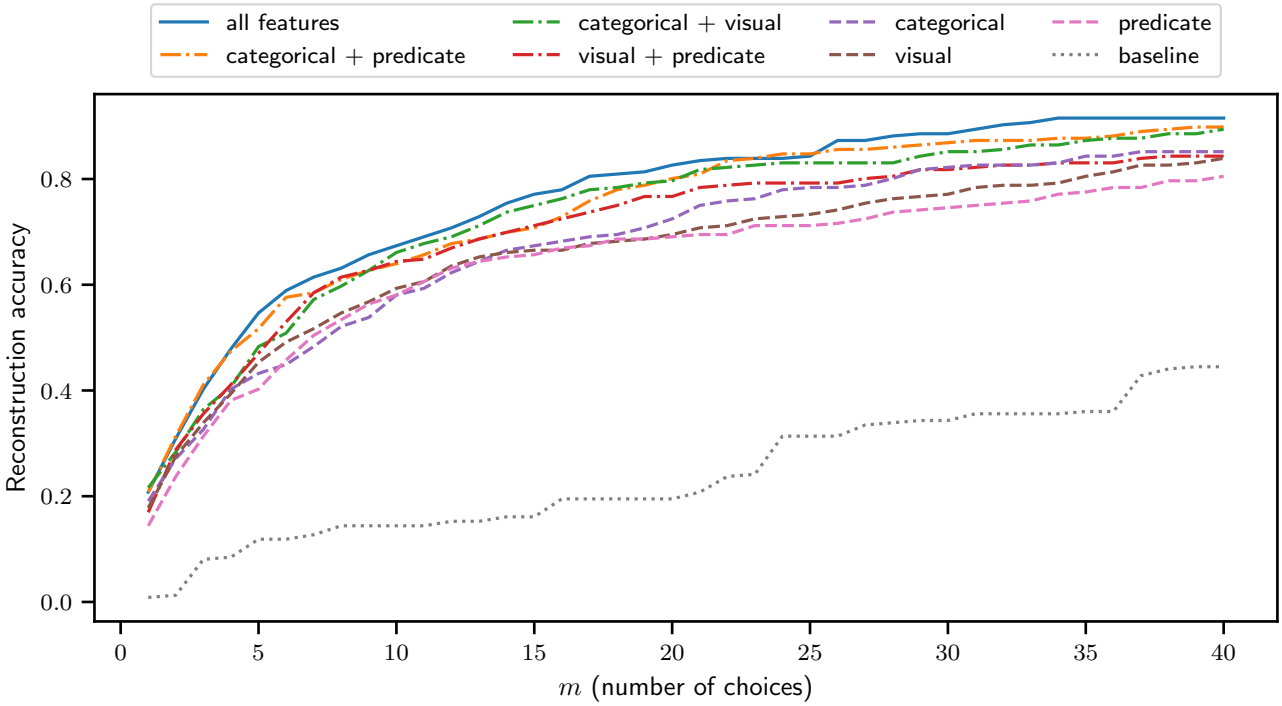

Figure S2: Performance curves for all production models showing cross-validated model accuracies in reconstructing overextended word choices ($N = 236$).

5.2 Sample predictions of the production model

Table S3 shows the top 5 words predicted by the production model for a random sample of the overextension dataset. Recalling that the top-5 model accuracy is approximately 55%, we observe that in the cases of incorrect model prediction, many of the predicted words are still closely related to the target concept, suggesting that even predictions that do not match the recorded child production are often sensible predictions. Since the overextension dataset is not an exhaustive record of child speech, this performance measure should be seen as a lower bound on the ability of the model to reconstruct children's overextension strategies.

Table S3: Top 5 model-predicted word choices for a random sample of the overextension dataset, stratified by correctness of model predictions. The upper panel shows examples for which the model predicts the true child production, and the lower panel shows examples for which model predictions do not include the child word. Each row displays the child production and intended referent from the overextension dataset, and the top 5 words predicted by the model, all denoted by their corresponding WordNet *synsets*.

Production	**Referent**	**1**	**2**	**3**	**4**	**5**
apple.n.01	banana.n.02	**apple.n.01**	orange.n.01	fruit.n.01	pea.n.02	juice.n.01
horse.n.01	deer.n.01	elk.n.01	animal.n.01	sheep.n.01	cow.n.01	**horse.n.01**
truck.n.01	train.n.01	car.n.01	bus.n.01	**truck.n.01**	bicycle.n.01	toy.n.03
sock.n.01	stocking.n.01	shoe.n.01	**sock.n.01**	hat.n.01	shirt.n.01	chair.n.01
dad.n.01	man.n.01	son.n.01	girl.n.01	baby.n.01	**dad.n.01**	animal.n.01
bicycle.n.01	motorcycle.n.01	car.n.01	**bicycle.n.01**	truck.n.01	wheel.n.01	train.n.01
catsup.n.01	mayonnaise.n.01	cheese.n.01	egg.n.02	peanut_butter.n.01	**catsup.n.01**	juice.n.01
ball.n.01	marble.n.02	**ball.n.01**	toy.n.03	chair.n.01	table.n.02	box.n.01
cheese.n.01	butter.n.01	**cheese.n.01**	milk.n.01	food.n.01	egg.n.02	toast.n.01
shoe.n.01	slipper.n.01	**shoe.n.01**	sock.n.01	blanket.n.01	hat.n.01	boot.n.01
kitten.n.01	horse.n.01	domestic_ass.n.01	pony.n.01	cow.n.01	animal.n.01	hog.n.03
cow.n.01	fish.n.01	tuna.n.03	animal.n.01	duck.n.01	child.n.01	baby.n.01
cat.n.01	lamb.n.01	sheep.n.01	animal.n.01	kitten.n.01	baby.n.01	dog.n.01
ball.n.01	peach.n.03	apple.n.01	orange.n.01	fruit.n.01	plum.n.02	grape.n.01
horse.n.01	jaguar.n.01	tiger.n.02	animal.n.01	lion.n.01	bear.n.01	cat.n.01
catsup.n.01	dressing.n.01	food.n.01	juice.n.01	cheese.n.01	pizza.n.01	pickle.n.01
bubble.n.01	marble.n.02	ball.n.01	toy.n.03	chair.n.01	table.n.02	box.n.01
horse.n.01	dog.n.01	puppy.n.01	animal.n.01	cat.n.01	kitten.n.01	kitty.n.04
cat.n.01	hog.n.03	cow.n.01	baby.n.01	animal.n.01	dog.n.01	bear.n.01
banana.n.02	tomato.n.01	apple.n.01	potato.n.01	juice.n.01	carrot.n.03	cheese.n.01

6 Production-comprehension results under an alternative comprehension model

Section 6.2 in the main text shows that our models replicate the observed patterns of overextension in production and comprehension, as well as the asymmetries between the two—namely, that children often overextend words in production that they understand correctly in comprehension. Since our production and comprehension models in Equations 1 and 4 share the same underlying multimodal semantic space and differ in the choice of prior, one could ask whether the predicted asymmetry is entirely due to the choice of a frequency prior for production and uniform prior for comprehension.

Here, we show that while this is not the case and that our main results also hold under an alternative comprehension model that uses a frequency prior for referents, there is at least one aspect of empirical data that our original formulation with different priors for production and comprehension captures better than the alternative model we present here. In other words, both the construction

of multimodal semantic space and choice of priors for production and comprehension seem to contribute to the best explanation of empirical data, supporting the hypotheses we proposed in the main text.

Concretely, we specify the following model of overextension in comprehension:

$$p_{\text{comp}}(c|w) = \frac{p(w|c)p(c)}{\sum_{c' \in E} p(w|c')p(c')} \tag{2}$$

The likelihood term $p(w|c)$ measures the appropriateness of word w to refer to concept c, and is defined by the same multimodal semantic similarity function as the other models: $p(w|c) = f_{\text{sim}}(c_w|c)$, where c_w is the concept corresponding to word w. Then, crucially, we define the prior $p(c)$ over referents to be proportional to their corresponding word frequencies in child-directed speech:

$$p(c) = \frac{F(w_c)}{\sum_{w' \in V} F(w')} \tag{3}$$

where w_c is the word corresponding to concept c, and $F(w_c)$ is the total frequency of word w_c in a representative corpus of children's linguistic environment.

Figure S3 shows the empirical data and model predictions under the same experimental settings as in Section 6.2. We observe that the model still captures the main empirical trends: performance in production is inferior for late nouns than for early nouns, and performance in production is overall inferior to performance in comprehension.

On the other hand, one aspect in which this model does not match empirical data as well as our original model is in the difference in comprehension performance between early and late nouns, which is predicted by the model but not shown in the empirical data; this discrepancy suggests that the role of word frequency is not as large in comprehension as in production, as we proposed in our original formulation. However, it is worth noting that while word frequency did not seem to affect comprehension in this particular lab study, it could still play a role in comprehension under different experimental settings; our framework opens the avenue for future empirical work to test this hypothesis rigorously. Nevertheless, by showing that our main results are robust to the choice of prior, we verified that our probabilistic formulation, together with the multimodal semantic space, capture patterns of young children's linguistic abilities in production and comprehension.

7 Validation of linguistic data from McDonough (2002)

The computational replication of McDonough [1] in Section 6.2 in the main text follows the original experiment in dividing the stimuli into early and late items (in age of acquisition). Figure S4 shows that the relative frequency data collected from child-directed speech reflects this division, i.e., on average, early nouns are more frequent than late nouns. This result verifies that our data is in sufficient agreement with [1] for a meaningful computational reproduction.

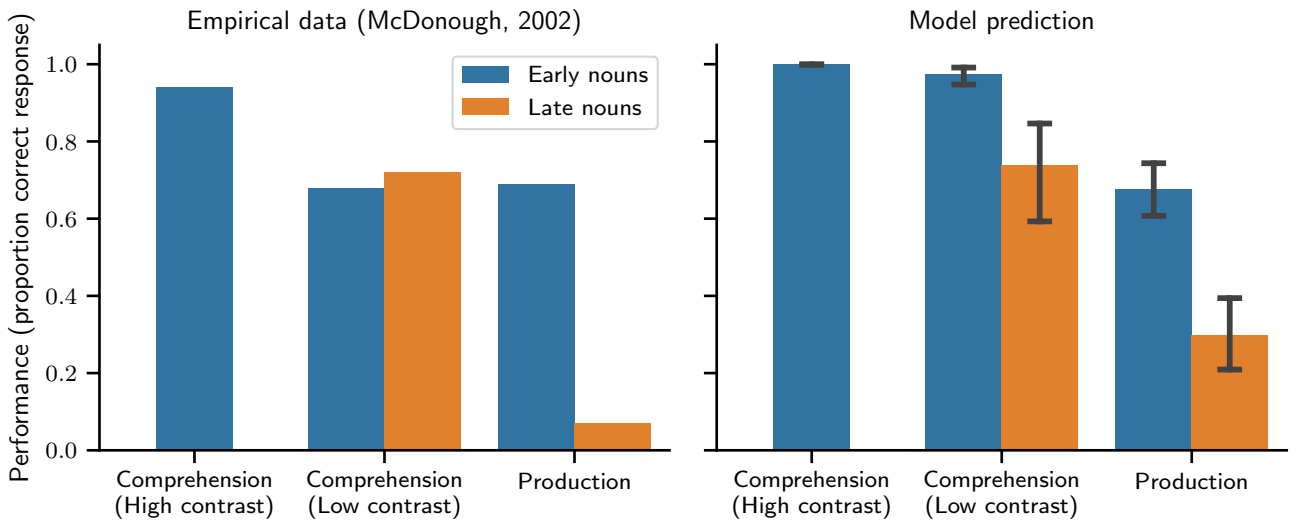

Figure S3: Results of comprehension and production experiments from empirical data of McDonough [1] and from reproduction under models of production and comprehension with frequency-based priors. Each bar shows the proportion of correct responses (referent selection in comprehension, and word utterance in production). Comprehension bars show performance over 14 triplets of stimuli, and production bars show performance over 16 early nouns and 14 late nouns. Error bars represent bootstrap 95% confidence intervals.

8 Vocabulary from early childhood

Table S4 shows the approximate vocabulary from early childhood extracted from Wordbank and used in our analyses, with each word manually coded as a WordNet [2] *synset* to enable its representation in the semantic space.

Table S4: Approximate vocabulary from early childhood. Each cell shows the WordNet *synset* corresponding to one word in the vocabulary.

Synset			
airplane.n.01	alligator.n.02	animal.n.01	ant.n.01
apple.n.01	applesauce.n.01	aunt.n.01	baby.n.01
baby_buggy.n.01	bag.n.04	ball.n.01	balloon.n.01
banana.n.02	basement.n.01	basket.n.01	bat.n.01
bath.n.01	bathroom.n.01	bathtub.n.01	beach.n.01
bean.n.01	bear.n.01	bed.n.01	bedroom.n.01
bee.n.01	beer.n.01	belt.n.02	bench.n.01
beverage.n.01	bicycle.n.01	bird.n.01	bite.n.04
black.n.01	blanket.n.01	block.n.03	blue.n.01
boat.n.01	book.n.01	boot.n.01	bottle.n.01

bowl.n.01	box.n.01	breakfast.n.01	broom.n.01
brush.n.02	bubble.n.01	bucket.n.01	bug.n.01
bulge.n.01	bunny.n.02	bus.n.01	business_district.n.01
butter.n.01	butterfly.n.01	button.n.01	cake.n.03
camera.n.01	camping.n.01	can.n.01	candy.n.01
car.n.01	carrot.n.03	cat.n.01	catsup.n.01
chair.n.01	chamberpot.n.01	cheese.n.01	chewing_gum.n.01
chicken.n.02	child.n.01	chip.n.04	chocolate.n.03
church.n.02	clock.n.01	cloud.n.02	clown.n.02
coat.n.01	coca_cola.n.01	cock.n.04	coffee.n.01
comb.n.01	corn.n.01	cow.n.01	cracker.n.01
crayon.n.01	crib.n.01	cup.n.01	cupboard.n.01
dad.n.01	dance.n.02	deer.n.01	diaper.n.01
dinner.n.01	dish.n.01	doctor.n.01	dog.n.01
doll.n.01	domestic_ass.n.01	door.n.01	doughnut.n.02
drawer.n.01	dress.n.01	dryer.n.01	duck.n.01
dwelling.n.01	egg.n.02	elephant.n.01	elk.n.01
face.n.01	fire_engine.n.01	fireman.n.04	fish.n.01
flag.n.01	flower.n.01	fly.n.01	food.n.01
foot.n.01	fork.n.01	french_fries.n.01	friend.n.01
frog.n.01	fruit.n.01	game.n.09	garage.n.01
garbage.n.03	garden.n.01	gelatin.n.02	giraffe.n.01
girl.n.01	glass.n.02	glove.n.02	goose.n.01
grandfather.n.01	grandma.n.01	grape.n.01	grass.n.01
green.n.01	gym_shoe.n.01	hair.n.06	hamburger.n.01
hammer.n.02	hand.n.01	hat.n.01	head.n.01
helicopter.n.01	hen.n.01	hog.n.03	horse.n.01
hose.n.03	house.n.01	ice.n.01	ice_cream.n.01
ice_lolly.n.01	jacket.n.01	jar.n.01	jean.n.01
jello.n.01	juice.n.01	key.n.01	kitchen.n.01
kitten.n.01	kitty.n.04	knife.n.01	lady.n.01
lamb.n.01	lamp.n.01	lawn_mower.n.01	light.n.02
lion.n.01	lip.n.02	living_room.n.01	lollipop.n.02
lunch.n.01	ma.n.01	man.n.01	melon.n.01
menagerie.n.02	milk.n.01	mitten.n.01	monkey.n.01
mother.n.01	motorcycle.n.01	mouse.n.01	mouth.n.01
movie.n.01	muffin.n.01	nail.n.02	napkin.n.01
necklace.n.01	nurse.n.01	nut.n.01	octopus.n.02
onion.n.01	orange.n.01	oven.n.01	owl.n.01
paint.n.01	pancake.n.01	paper.n.01	party.n.02
patty.n.01	pea.n.02	peach.n.03	peanut_butter.n.01
pen.n.01	pencil.n.01	penguin.n.01	people.n.01
person.n.01	pickle.n.01	picnic.n.03	pillow.n.01
pizza.n.01	plant.n.02	plate.n.04	playground.n.02
plum.n.02	pony.n.01	pop.n.02	popcorn.n.01

porch.n.01	potato.n.01	pretzel.n.01	pudding.n.01
pumpkin.n.01	puppy.n.01	puzzle.n.02	radio.n.01
raisin.n.01	rear.n.05	refrigerator.n.01	rock.n.01
rocking_chair.n.01	roof.n.01	room.n.01	salt.n.02
sandwich.n.01	sauce.n.01	scarf.n.01	school.n.02
scissors.n.01	sheep.n.01	shirt.n.01	shoe.n.01
shop.n.01	short_pants.n.01	shoulder.n.01	shovel.n.01
shower.n.01	sidewalk.n.01	sink.n.01	sister.n.01
skate.n.01	sky.n.01	sled.n.01	slide.n.04
slide_fastener.n.01	slipper.n.01	snow.n.01	sock.n.01
sofa.n.01	son.n.01	soup.n.01	spaghetti.n.01
spectacles.n.01	spoon.n.01	sprinkler.n.01	squirrel.n.01
stairs.n.01	star.n.03	stick.n.01	stove.n.01
strawberry.n.01	street.n.01	sun.n.01	swab.n.02
sweater.n.01	swing.n.02	table.n.02	tape.n.04
tea.n.01	teacher.n.01	telephone.n.01	television.n.01
tiger.n.02	tights.n.01	toast.n.01	tooth.n.02
toothbrush.n.01	towel.n.01	toy.n.03	tractor.n.01
train.n.01	tray.n.01	tree.n.01	tricycle.n.01
trouser.n.01	truck.n.01	tuna.n.03	turkey.n.01
turtle.n.02	underpants.n.01	vacuum.n.04	vanilla.n.01
vitamin.n.01	walker.n.04	wash.n.01	washer.n.03
watch.n.01	water.n.06	wheel.n.01	white.n.02
window.n.01	wolf.n.01	yellow.n.01	yogurt.n.01
zebra.n.01			

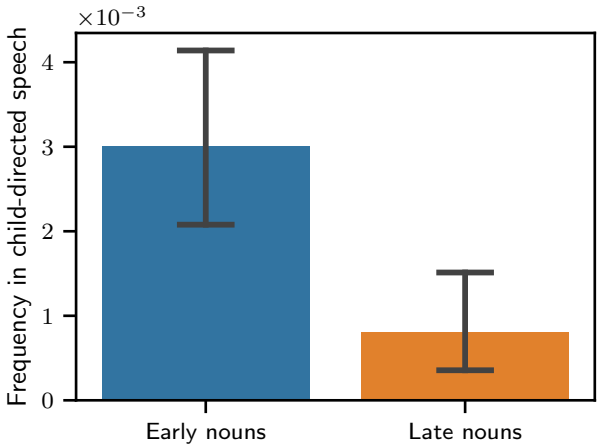

Figure S4: Relative frequencies of 16 early and 14 late nouns from McDonough [1] in child-directed speech. Error bars represent bootstrap 95% confidence intervals.

References

[1] Laraine McDonough. Basic-level nouns: first learned but misunderstood. *Journal of Child Language*, 29(2):357–377, 2002.

[2] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38 (11):39–41, 1995.

[3] Leslie A. Rescorla. Category development in early language. *Journal of Child Language*, 8 (2):225–238, 1981.