Predicting children's early word learning using egocentric videos of their learning environments

Alvin W. M. Tan[1] **(tanawm@stanford.edu)**
Khai Loong Aw[1] **Tarun Sepuri**[2] **Robert Z. Sparks**[1] **Jane Yang**[2]
Virginia A. Marchman[1] **Bria Long**[2] **Michael C. Frank**[1]
[1]Department of Psychology, Stanford University
[2]Department of Psychology, University of California San Diego

Abstract

How do features of children's environmental input predict their word learning? Prior research has suggested that distributional features of the language children hear (e.g., frequency and syntactic complexity) predict variation in early word learning, but these studies have typically relied on aggregated estimates from disjunct samples to calculate input features and word learning outcomes. In this study, we use child-specific input features derived from 1118h of at-home egocentric video recordings to show that word frequency, length in phonemes, and syntactic complexity predict word knowledge within individual children (N = 29). The predictive power of these within-child features was greater than features aggregated across children, supporting theories positing a direct relationship between input distributions and children's language learning. Exploratory analyses involving object frequency, word–object cooccurrences, and context distinctiveness did not reveal any additional effects of these predictors. The use of child-specific naturalistic data provides an avenue for more precise and comprehensive investigations of input–outcome relationships for word learning, allowing researchers to adjudicate among theories of language learning in young children.

Keywords: word learning; input; frequency; context

Introduction

Young children exhibit remarkable variability in their language abilities (Fenson et al., 1994; Frank et al., 2021). An important source of such individual differences is the variability in children's environmental input. Prior research has suggested that the quantity and quality of caregiver input influences children's language development; in particular, children who receive more, more diverse, and more complex language input have larger vocabularies (Anderson et al., 2021; Cartmill et al., 2013; Cychosz et al., 2021; Cychosz et al., 2025; Hoff, 2003; Huttenlocher et al., 1991, 2010; Newman et al., 2016; Rowe, 2012; Weizman & Snow, 2001).

What are the causal mechanisms that relate input features to language outcomes? Some previous efforts have sought to constrain possible mechanisms of language learning by zooming in to the level of words—examining how word features relate to the likelihood that they are produced by children (Frank et al., 2021; Tan et al., 2024). For example, words that are more frequent are more likely to be learnt earlier than words that are less frequent (e.g., Goodman et al., 2008), lending support to accumulator models of word learning (e.g., Kachergis et al., 2022), whereby repeated exposures to a word accumulate over time, until sufficient information is accrued and the word becomes known.

However, one major challenge of these approaches has been that the input features were estimated over a different set of children than the outcome measures—for example, Goodman et al. (2008) estimated word frequency from speech copora in CHILDES (MacWhinney, 2000), whereas the age of acquisition of words was estimated using data from the MacArthur-Bates Communicative Development Inventory (CDI; Marchman et al., 2023) collected on a different set of children. One exception is Huttenlocher et al. (1991), which used frequency estimates from audio recordings of mother–child interactions to predict the ages of acquisition of content words for those children. Notably, they found that the relative frequencies of words were highly correlated across children, and thus collapsed results across children, obtaining the overall correlation between average log word frequency and average word age of acquisition, rather than doing individual-level prediction.

Moving toward individual-level prediction of word learning trajectories would afford tighter constraints on learning mechanisms, demonstrating that it is specifically features of the child's own input that relate to their word learning, rather than general input-independent features of a word. This approach would also bring us closer to a "precision science" of language learning, permitting individualised predictions and targeted interventions (Samuelson, 2021). Additionally, language learning does not occur in a speech-only vacuum, but in rich multimodal environments, with multiple sources of information jointly contributing to learning (Bohn & Casillas, 2026; Kosie & Lew-Williams, 2024). Fortunately, recent advances in data collection methods and datasets (e.g., Bergelson et al., 2019; Geangu et al., 2023; Long et al., 2023) as well as advances in data processing tools (e.g., Bang et al., 2023; Sepuri et al., 2025) have allowed researchers to study features of children's environmental input beyond merely the language stream itself, including wider-ranging visual and contextual features which may support word learning (Cain et al., 2025; Kachergis et al., 2017; Roy et al., 2015).

Understanding such features is crucial for the development of theories of grounded language learning. One such theory is cross-situational word learning, which suggests that children learn word–object correspondences by accumulating many instances of seeing a referent in conjunction with its label, helping to disambiguate among possible word–object relations (Cain et al., 2025; Smith & Yu, 2008; Suanda et

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

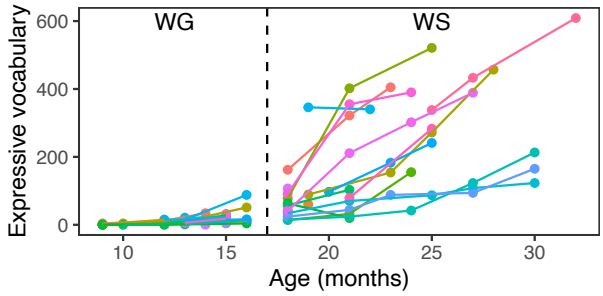

Figure 1: Total expressive vocabulary of BabyView participants by age. Colours represent individual participants. Dashed line indicates whether the WG or the WS form was administered.

al., 2014). Under this theory, word–object cooccurrence frequency should predict word learning—words that consistently occur with their referents are more likely to be learnt. Modern computational models that learn from naturalistic vision–language cooccurrences (e.g., Vong et al., 2024; S. Wang et al., 2025) also lend support to the plausibility of cross-situational word learning as a language learning mechanism.

Another theory of word learning focuses on the role of context, which may structure children's expectations about the kinds of social interactions, visual experiences, and language distributions they receive within any given situation (Bang et al., 2025; Goldenberg et al., 2022; Sepuri et al., 2025). Thus, words which occur in more distinctive contexts may have meanings which are more easily inferred, and therefore may be learnt earlier (Roy et al., 2015). Both of these theories remain to be validated using large, multi-child, naturalistic data to determine the extent to which children *actually* employ such mechanisms in learning words.

In this paper, we analysed the BabyView dataset (Long et al., 2025), a recent large-scale dataset of young children's at-home egocentric video recordings, to understand how environmental input relates to vocabulary outcomes. We did so across a set of four complementary analyses. First, we did (1) a replication of prior work showing that overall characteristics of caregiver speech input are related to overall vocabulary ability, and (2) an item-wise analysis showing that linguistic predictors of word knowledge are most predictive when estimated for a specific timepoint for a specific child, rather than aggregating data. We then exploratorily investigated (3) possible visual and vision–language predictors as well as (4) the role of context effects on word knowledge, finding no effects unlike in previous work. Together, these analyses represent an opportunity to study complex, multimodal input to young children, and provide a step towards developing more robust and well-specified models of word learning mechanisms.

Analysis 1: Overall vocabulary

Previous studies have demonstrated that features of caregiver speech predict young children's language abilities (e.g., Cartmill et al., 2013; Huttenlocher et al., 1991). Here, we replicate these results on the children in the BabyView dataset, examining the roles of input quantity, lexical diversity, and syntactic complexity in predicting vocabulary ability.

Methods

Dataset. For this and all subsequent analyses, we analyse data from the BabyView dataset (Long et al., 2023, 2025), a set of naturalistic at-home videos obtained from a head-mounted GoPro Hero Bones camera worn by infants and young children. We use data from the 2025.2 release of the dataset, which includes 1428 hours of videos from 37 families; we only included data from children who were exposed to ≥80% English (1118h of videos from 29 children). These videos are accompanied by expressive vocabulary data collected approximately every 3 months using the American English CDI, which we use as our outcome data. The Words & Gestures (WG) form was used with children aged 16 months and younger, while the Words & Sentences (WS) form was used with children aged 17 months and older. In total, we included 42 WG administrations and 52 WS administrations.

Transcripts. To examine the role of language input features, we first transcribed the audio from the videos using WhisperX, specifically the `large-v3` model (Bain et al., 2023; Radford et al., 2022). The model also provided word-level timestamps. We then used VTC 2.0 (Charlot et al., 2025) to identify the voice type of each utterance (male adult, female adult, key child, or other child). Since we were interested in the role of input, we filtered the transcripts down to adult-produced utterances for all analyses; adult speech also had the most accurate transcripts and voice type classification (Long et al., 2025). We lemmatised the transcripts using spaCy with the `en_core_web_sm` model (Honnibal et al., 2023).[1]

Vocabularies. The raw expressive vocabulary data from the CDIs are shown in Figure 1. The longitudinal nature of the CDI presented two challenges. First, the CDI is bounded, and exhibits floor and ceiling effects at younger/older ages respectively. Second, the WG and WS forms had partially overlapping but non-identical item sets. To address these issues, we adopted an item response theory (IRT) approach, which simultaneously models the difficulty of each CDI item and the vocabulary ability of each child at each administration (Embretson & Reise, 2000). We first stitched the WG and WS item sets by combining items that were identical across both forms (see Kachergis et al., 2025), then estimated vocabulary abilities using a 2-parameter logistic model previously fit to the American English production data from Wordbank (Frank et al., 2017; see Kachergis et al., 2022). The model estimated vocabulary abilities for each child at each administration on a

[1] For pronouns and auxiliary verbs, we preserved the original word form instead of using the lemma, since we expect that these frequent closed class words may have distinct learning trajectories, such that collapsing them into a single lemma may not be appropriate. This approach also better afforded subsequent comparisons with the CDI, since it includes several forms of pronouns and auxiliary verbs as different items.

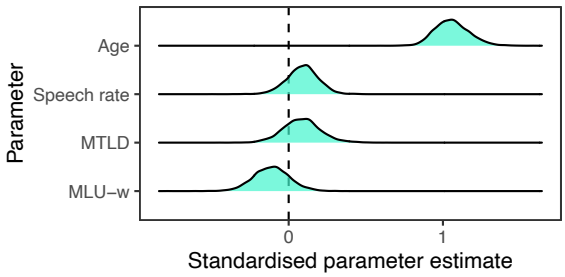

Figure 2: Posterior distributions of predictor coefficient estimates from the vocabulary ability model. Dashed line indicates zero.

continuous scale that was approximately linear in age, which we used as the outcome variable in this analysis.

Metrics. We calculated three language input metrics from the transcripts for each child: the *adult speech rate* (in words per minute) as a measure of input quantity, the *measure of textual lexical diversity* (MTLD; McCarthy & Jarvis, 2010) as a measure of lexical diversity, and the *mean length of utterance in words* (MLU-w) as a measure of syntactic complexity. All metrics were standardised (0-centred and 1-scaled).

Analysis. We constructed a Bayesian model predicting estimated vocabulary ability with the fixed effects of adult speech rate, MTLD, MLU-w, and standardised age, as well as random intercepts and age slopes for each child. We used weakly informative priors of $t(3,0,2)$ for coefficients and Cauchy(0,2) for standard deviations of random effects for all analyses except where otherwise noted. We determined the reliability of effects by examining whether the 95% credible intervals (CrIs) overlapped with zero.

Results and discussion

Figure 2 shows the posterior distributions of the fixed effects. The only reliable effect was the expected effect of age (β = 1.05 [0.86, 1.28]), indicating that older children had greater vocabulary abilities. While we did not find reliable effects of the input features, the posterior estimates lay in the expected directions, such that children who received more speech (β = 0.11 [-0.12, 0.28]), more lexically diverse speech (β = 0.11 [-0.17, 0.32]), and less syntactically complex speech (β = -0.10 [-0.32, 0.14]) tended to have greater vocabulary abilities; the probability of direction for these effects were p_d = 0.823 (speech rate), 0.763 (MTLD), and 0.835 (MLU-w) respectively. These results partially replicated prior findings on the role of input features on young children's vocabulary; the lack of reliable effects may have been due to the relatively small sample size in this analysis.

Analysis 2: Item-level prediction

To zoom in on more specific relationships between environmental input and word learning, we next turned to an item-level approach to examine how input features related to whether children produced specific words on the CDI. This

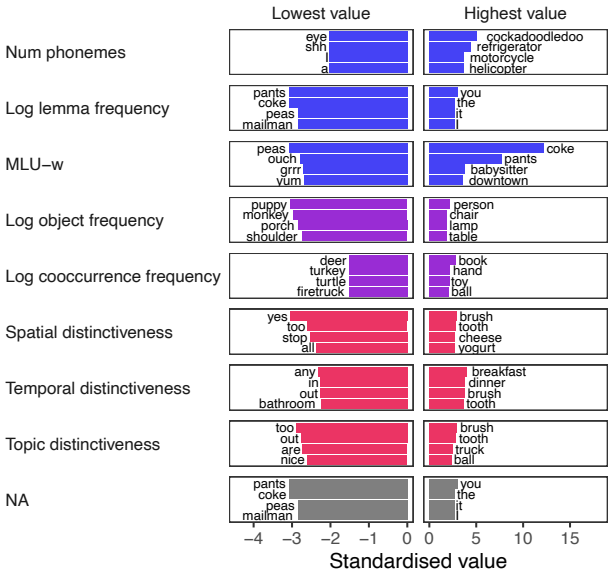

Figure 3: Items with the lowest and highest standardised values across all predictors used in Analyses 2–4.

approach allowed us to investigate within-child variability across items, and would also help to increase the power of our analyses by increasing the number of observations. Here, we sought to replicate prior work showing that key word-level input features (frequency, MLU-w, and number of phonemes) predicted whether children were able to produce the word (Frank et al., 2021; Goodman et al., 2008; Tan et al., 2024).

We additionally investigated the differences between three methods of estimating these input features: using data aggregated across all children and videos (*global* metrics), data aggregated separately for each child across all videos for that child (*by-child* metrics), or data only from videos prior to the CDI administration for each child (*local* metrics). These three levels reflected a bias–variance trade-off: global metrics were more likely to be precise, minimising the impact of noisy variation across videos, but they would fail to capture the variation in input across families and across time. Conversely, local metrics would have lower bias since they specifically captured the relevant input that would contribute to word learning in any given child at any given timepoint, but metric estimates might be noisier. Local metrics would also best support accumulator models of word learning (Kachergis et al., 2022), since word knowledge is directly predicted from preceding instances of the word.

Methods

Instead of using the IRT-estimated vocabulary abilities as in Analysis 1, we used the raw CDI item-level production data as our outcome measure. Using the lemmatised transcripts, we calculated the lemma frequency of each CDI lemma by counting the number of occurrences, applying a Laplace smoothing (adding 1 to all counts), then expressing frequency as smoothed token count divided by total video du-

Table 1: ELPD comparisons across levels of estimation for item-level linguistic models.

		ΔELPD from best model	
Analysis	Model	Estimate	95% CI
Analysis 2	Baseline	-199.87	[-245.52, -154.23]
	Global	-198.04	[-244.38, -151.70]
	By-child	-10.36	[-29.33, 8.62]
	Local	**0.00**	—
Analysis 3	Linguistic	-1.58	[-2.77, -0.40]
	+ log object freq	**0.00**	—
	+ log cooc freq	-3.06	[-4.22, -1.90]
	Full	-0.55	[-1.53, 0.44]
	NA	-1.39	[-2.16, -0.63]
Analysis 4	Linguistic	-17.64	[-33.98, -1.30]
	+ spatial dist	-21.25	[-36.75, -5.74]
	+ temporal dist	-9.62	[-22.10, 2.85]
	+ topic dist	-1.46	[-12.73, 9.80]
	Full	-1.29	[-7.40, 4.83]
	NA	**0.00**	—

ration, which was then log-transformed. We also calculated the mean length of utterance in words (MLU-w) for all CDI lemmas. Lemma frequency and MLU-w were calculated at three levels: global, by-child, and local, as described above. We calculated the number of phonemes for each CDI lemma using the CMU Pronouncing Dictionary (Carnegie Mellon Speech Group, 2026) to obtain phonemic transcriptions for each lemma, and counting the number of phonemes in each transcription. Again, all metrics were standardised.

Figure 3 shows lemmas with high and low values for each metric (using the global metrics), including all metrics used in Analyses 2–4.

Analysis. We constructed four models. The first was a baseline growth curve model, with age as a fixed effect, as well as random intercepts and age slopes for each child and random intercepts for each CDI item. The remaining models added the fixed effects of lemma frequency, MLU-w, and number of phonemes; the models respectively used global, by-child, and local estimates, and also had interactions with lexical categories for these item-level predictors.

We then measured the estimated log pointwise predictive density (ELPD) for each model using approximate leave-one-out cross validation. The model with greatest ELPD was considered to best fit the data, and models with ≥4 ELPD less than the best model were considered to be substantially worse than the best model (Sivula et al., 2022).

Results and discussion

Table 1 shows the ELPD differences between models, with the best model as the zero point. The local metrics model was best, and although no model was within 4 ELPD from the best model, the by-child metrics model had an ELPD confidence interval that included zero (i.e., not reliably different from the best model). Both the local and by-child metrics outperformed the global metrics, suggesting that the specific linguistic input distributions received by the child were more tightly related to word learning than distributional properties of the words in general. Thus, the amount of data in the 2025.2 release of BabyView was sufficient to result in the best estimates from the local metrics model; hence, we use local metrics for all subsequent analyses. The posterior estimates from the local metrics model are displayed in Figure 4A, showing that words that are more frequent and less syntactically complex are generally more likely to be known, with some variation by lexical category. The effect of number of phonemes was not reliably different than zero for most lexical categories. Further simulation studies can help to investigating the role of data quantity in determining what level of estimation would be the most effective for any given amount of data.

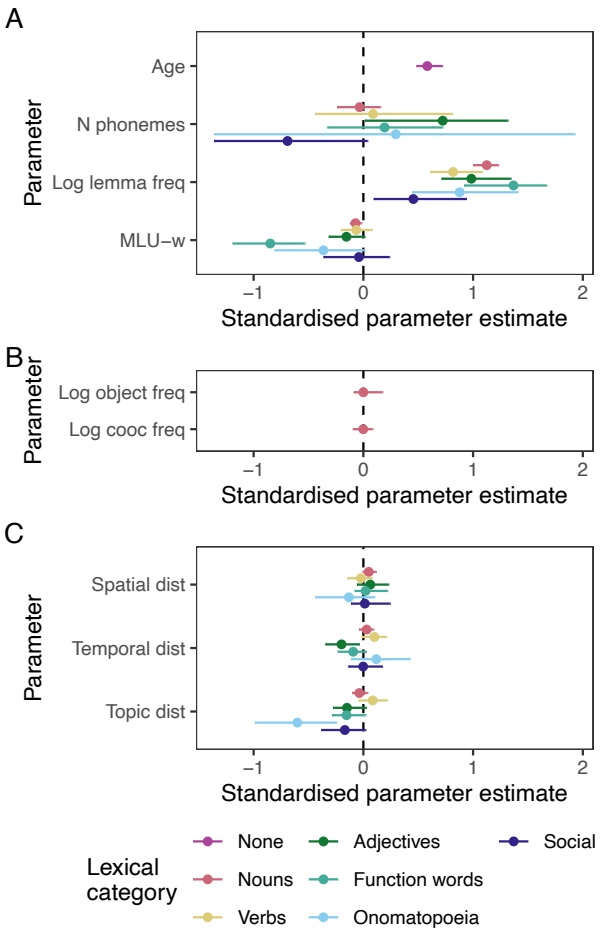

Figure 4: (A) Predictor coefficient estimates from the local metrics linguistic model (Analysis 2). (B) Coefficient estimates of object and cooccurrence frequencies from the full visual / cross-modal model (Analysis 3). (C) Coefficient estimates of distinctiveness predictors from the full distinctiveness model (Analysis 4). Dashed line indicates zero. Error bars indicate 95% CrIs.

Analysis 3: Visual and cross-modal predictors

In this analysis, we examined the possible roles of object frequency and label–object cooccurrences in word learning.

Methods

Annotation. We identified objects appearing in frame by extracting frames at 1 fps, then conducting object detection using YOLO-E (A. Wang et al., 2025). We used the concrete nouns present on the CDI as the vocabulary set for object detection (Yang et al., 2025).

Metrics. We estimated object frequency by counting the number of occurrences, applying a Laplace smoothing, then dividing smoothed object count by total video duration and log-transforming. We estimated the number of object–word cooccurrences by determining whether an object was present within a ±5s window around the utterance of the corresponding lemma. The number of cooccurrences was converted into a frequency by smoothing, dividing by the total video duration, and log-transforming.[2] Object statistics were only calculable for concrete nouns.

Analysis. Preliminary descriptive statistics revealed moderate correlations among some predictors; hence, we adopted a two-pronged approach to determining the possible contribution of cross-situational word learning predictors. First, we constructed models that contained the linguistic predictors from Analysis 2 (number of phonemes, log lemma frequency, and MLU-w) along with *either* log object frequency or log cooccurrence frequency individually. Second, we constructed a full model that contained the linguistic predictors as well as *both* log object and cooccurrence frequencies, using horseshoe priors (with df = 3) for regularisation (Piironen & Vehtari, 2017). We also included a model with only linguistic predictors as a baseline. Note that these models did not include lexical category as a predictor, since all items were nouns. We again estimated the ELPD for each model for model comparison.

Results and discussion

Table 1 shows the ELPD differences between models, with the best model as the zero point. The model including log object frequency is the best model, although all models were within 4 ELPD of the best model; thus, they are not substantially different than each other in terms of predictive performance. For brevity, we display only the parameter estimates of the object and cooccurrence predictors for the full model in Figure 4B. Both log object frequency and log cooccurrence frequency effectively did not predict word knowledge. These results suggest that naïve cooccurrences are insufficient for word learning, even when only considering concrete nouns; additional supporting information including speakers' referential intention (Frank et al., 2009; Jordan-Barros et al., 2025), contingent behaviour in social interaction (Bianco et al., 2025), or the sizes of objects in the visual field (Cain et al., 2025) may help to increase the likelihood of learning from valid word–object pairings while ignoring spurious pairings.

[2]Note that pointwise mutual information is expressible as a linear combination of log cooccurrence frequency, log lemma frequency, and log object frequency; thus, a model including these three terms would also represent the joint informativeness of the lemma and object signal streams.

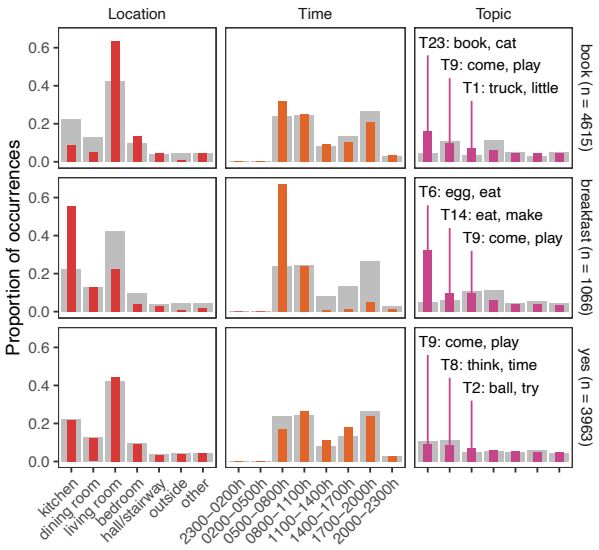

Figure 5: Location, time, and topic distributions for 3 sample words ("book", "breakfast", "yes"). For visualisation purposes, locations and times are binned into broader categories, and only the top 7 topics for each word are shown. Labelled topics reflect the 2 most frequent lemmas for the top 3 topics for each word (after filtering out the 15 most frequent lemmas overall). Grey bars reflect average distributions.

Analysis 4: Context distinctiveness

In this analysis, we replicate the analyses in Roy et al. (2015) to determine the robustness of context distinctiveness effects across datasets and children.

Methods

Annotation. We determined the location of the child wearing the camera (e.g., kitchen, bedroom, garage) by splitting videos into 10-second clips, then prompting VideoLLaMA 3 (Zhang et al., 2025) to identify the location of each clip out of a manually constructed list of options (Sepuri et al., 2025).

Metrics. We calculated the spatial, temporal, and topic distinctiveness of all CDI lemmas using a similar method to Roy et al. (2015). The distinctiveness of a lemma along a feature reflects the extent to which that lemma is strongly associated with particular feature values, reflecting how specialised the contexts in which a lemma occurs are. Distinctiveness was determined by calculating the average distribution of that feature across all lemmas, measuring the Kullback–Leibler divergence of the distribution for one lemma from the average distribution, and residualising log divergence against log frequency. The residuals were then used as the distinctiveness values. Spatial divergence was estimated using the location

annotations. Temporal divergence was estimated by splitting videos into hourly time bins. Topic divergence was estimated by conducting latent Dirichlet allocation on the transcripts with $k = 25$ topics after removing stopwords and infrequent words ($\leq$5 instances or $\leq$4 transcripts), using an α parameter of 0.1. Figure 5 displays the spatial, temporal, and topic distributions for a few sample lemmas, compared with the average distribution across all lemmas.

Analysis. As in Analysis 3, we constructed models that contained linguistic predictors along with one of the distinctiveness metrics, as well as a full model that contained linguistic predictors and all distinctiveness metrics with horseshoe priors, and a model with only linguistic predictors as a baseline. We additionally included interactions with lexical category for all item-level predictors. We again estimated the ELPD for each model for model comparison.

Results and discussion

Table 1 shows the ELPD differences between models, with the best model as the zero point. The topic distinctiveness model is the best, although the full model had an ELPD confidence interval that included zero. For brevity, we display only the parameter estimates of the distinctiveness predictors for the full model in Figure 4C. All distinctiveness predictors broadly did not show reliable effects across lexical categories, except for negative effects for topic distinctiveness in adjectives and onomatopoeia. These results contrast from those of Roy et al. (2015), where the distinctiveness effects were positive and in fact dominated over the effects of lemma frequency.

Why do we observe different results? One crucial reason may be that the data from Roy et al. (2015) were collected over >70% of the child's waking hours, across most of the rooms of the family's house, whereas the caregivers of the children in the BabyView dataset recorded ~1h a week and were able to select the contexts during which they were recording; thus, the coverage of contexts was not as randomly and comprehensive selected, which may have led to biased estimates of context distinctiveness. Additionally, Roy et al. (2015) investigated the role of distinctiveness in predicting age of first production, which requires clear, decontextualised production decodable by an external observer, in contrast with the parent-report nature of the CDI.

General discussion

What are the features of children's environmental input that support word learning? Across four analyses, we found that a diverse set of input features relate to children's vocabulary knowledge, weakly at the overall ability level and more clearly at the item level. Words which were more frequent and less syntactically complex were more likely to be known, but we did not find any reliable effect of object frequency, cooccurrence frequency, or context distinctiveness (even though distinctiveness did improve predictive power). These results strengthen previous findings about input-dependent language learning, but also raise new questions about mechanisms of word learning.

Using input features estimated for specific children best predicted their own outcomes. The child-wise approach provides stronger evidence for mechanisms of word learning that connect children's input to learning, including accumulator models of word learning (e.g., Kachergis et al., 2022).

The null effects of object frequency, cooccurrence frequency, and distinctiveness metrics in Analyses 3 and 4 do not invalidate the theories that these analyses were inspired by; rather, they suggest that such mechanisms may be more nuanced and may operate differently across contexts and individuals, such that we were not able to capture their effect using the dataset, metrics, and statistical approach that we employed. It is possible that their effect sizes are small, and more data is needed to identify a reliable effect. Children may also use different mechanisms to differing extents over development, such that more age- or ability-sensitive models may be needed to identify their effects. Nonetheless, our analyses provide a proof-of-concept for the testing of word learning mechanisms at the by-child, item-wise level.

Some fundamental theoretical questions remain in establishing true mechanistic explanations of word learning. First, due to the correlations among input features (both observed and unobserved), it is possible that the features we selected were not the proximal features that directly influence word learning, but rather correlates of other (possibly unobserved) features that were causally involved. For example, children's motivational states are a key proximal factor in word learning that may also affect what caregivers choose to talk about and interact with—which may in turn affect lemma and object distributions. The use of multiple related features helps to mitigate this issue, but much more work is required to provide a comprehensive, multifaceted description of children's learning environments so as to better understand the regularities and relationships among different features of their input.

Second, the relationships between input features and outcomes may be consistent with multiple mechanistic pathways of word learning. This issue can similarly be addressed by investigating other predictors and outcomes, which may disambiguate predictions of different word learning theories. Using statistics from naturalistic data can also help to clarify the distinction between learning mechanisms that children can *possibly* use (which can be verified using in-lab experiments), and learning mechanisms that children *do actually* use in real-world learning scenarios (see e.g., Casey et al., 2023).

Conceptually, the investigation of specific relationships between environmental input and learning outcomes can be construed as a kind of dimensionality reduction, projecting from rich and high-dimensional input streams to meaningful dimensions of variation that predict learning. As datasets and data processing capacities continue to develop, we anticipate more thoughtful and complex analyses of naturalistic input to children, which will help to illuminate and adjudicate between theories of early learning, pushing our understanding of language development ever closer to the trajectories that children actually undertake as they master their language.

References

Anderson, N. J., Graham, S. A., Prime, H., Jenkins, J. M., & Madigan, S. (2021). Linking quality and quantity of parental linguistic input to child language skills: A meta-analysis. *Child Development*, *92*(2), 484–501. https://doi.org/10.1111/cdev.13508

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *INTERSPEECH 2023*, 4489–4493. https://doi.org/10.21437/Interspeech.2023-78

Bang, J. Y., Kachergis, G., Weisleder, A., & Marchman, V. (2023). An automated classifier for periods of sleep and target-child-directed speech from LENA recordings. *Language Development Research*, *3*(1). https://doi.org/10.34842/xmrq-er43

Bang, J. Y., Mora, A., Munévar, M., Fernald, A., & Marchman, V. A. (2025). Time to Talk: Variability in Caregiver-Child Verbal Engagement During Everyday Activities Sampled From Daylong Recordings. *Infancy*, *30*(6), e70051. https://doi.org/10.1111/infa.70051

Bergelson, E., Amatuni, A., Dailey, S., Koorathota, S., & Tor, S. (2019). Day by day, hour by hour: Naturalistic language input to infants. *Developmental Science*, *22*(1), e12715. https://doi.org/10.1111/desc.12715

Bianco, C., Pang, J., & Yu, C. (2025). Contingent behavior during caregiver-child interaction improves the quality of word learning opportunities. *2025 IEEE International Conference on Development and Learning (ICDL)*, 1–6. https://doi.org/10.1109/ICDL63968.2025.11204427

Bohn, M., & Casillas, M. (2026). Language learning as ontogenetic adaptation. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2025.12.005

Cain, E. S., Ryskin, R. A., & Yu, C. (2025). Cross-Situational Statistics Present in an Early Language Learning Context: Evidence From Naturalistic Parent–Child Interactions. *Cognitive Science*, *49*(6), e70078. https://doi.org/10.1111/cogs.70078

Carnegie Mellon Speech Group. (2026). *CMU Pronouncing Dictionary*. CMU Sphinx.

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, *110*(28), 11278–11283. https://doi.org/10.1073/pnas.1309518110

Casey, K., Potter, C. E., Lew-Williams, C., & Wojcik, E. H. (2023). Moving beyond "nouns in the lab": Using naturalistic data to understand why infants' first words include uh-oh and hi. *Developmental Psychology*, *59*(11), 2162–2173. https://doi.org/10.1037/dev0001630

Charlot, T., Kunze, T., Poli, M., Cristia, A., Dupoux, E., & Lavechin, M. (2025). *BabyHuBERT: Multilingual Self-Supervised Learning for Segmenting Speakers in Child-Centered Long-Form Recordings* (No. arXiv:2509.15001). arXiv. https://doi.org/10.48550/arXiv.2509.15001

Cychosz, M., Edwards, J. R., Bernstein Ratner, N., Torrington Eaton, C., & Newman, R. S. (2021). Acoustic-Lexical Characteristics of Child-Directed Speech Between 7 and 24 Months and Their Impact on Toddlers' Phonological Processing. *Frontiers in Psychology*, *12*, 712647. https://doi.org/10.3389/fpsyg.2021.712647

Cychosz, M., Romeo, R. R., Edwards, J. R., & Newman, R. S. (2025). Bursty, Irregular Speech Input to Children Predicts Vocabulary Size. *Developmental Science*, *28*(1), e13590. https://doi.org/10.1111/desc.13590

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Psychology Press. https://doi.org/10.4324/9781410605269

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., Tomasello, M., Mervis, C. B., & Stiles, J. (1994). Variability in Early Communicative Development. *Monographs of the Society for Research in Child Development*, *59*(5), i–185. https://doi.org/10.2307/1166093

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694. https://doi.org/10.1017/S0305000916000209

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and Consistency in Early Language Learning: The Wordbank Project*. MIT Press.

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. https://doi.org/10.1111/j.1467-9280.2009.02335.x

Geangu, E., Smith, W. A. P., Mason, H. T., Martinez-Cedillo, A. P., Hunter, D., Knight, M. I., Liang, H., Del Carmen Garcia de Soria Bazan, M., Tse, Z. T. H., Rowland, T., Corpuz, D., Hunter, J., Singh, N., Vuong, Q. C., Abdelgayed, M. R. S., Mullineaux, D. R., Smith, S., & Muller, B. R. (2023). EgoActive: Integrated Wireless Wearable Sensors for Capturing Infant Egocentric Auditory-Visual Statistics and Autonomic Nervous System Function 'in the Wild'. *Sensors (Basel, Switzerland)*, *23*(18), 7930. https://doi.org/10.3390/s23187930

Goldenberg, E. R., Repetti, R. L., & Sandhofer, C. M. (2022). Contextual variation in language input to children: A naturalistic approach. *Developmental Psychology*. https://doi.org/10.1037/dev0001345

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(3), 515–531. https://doi.org/10.1017/S0305000907008641

Hoff, E. (2003). The Specificity of Environmental Influence: Socioeconomic Status Affects Early Vocabulary Development Via Maternal Speech. *Child Development*, *74*(5),

1368–1378. https://doi.org/10.1111/1467-8624.00612

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2023). *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo. https://doi.org/10.5281/ZENODO.1212303

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, *27*(2), 236–248. https://doi.org/10.1037/0012-1649.27.2.236

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of Variability in Children's Language Growth. *Cognitive Psychology*, *61*(4), 343–365. https://doi.org/10.1016/j.cogpsych.2010.08.002

Jordan-Barros, A., Cabiddu, F., Donnellan, E., Gu, Y., & Vigliocco, G. (2025). *Caregivers' multimodal actions scaffold word learning and vocabulary growth in the early years* (No. bdnu2_v1). PsyArXiv. https://doi.org/10.31234/osf.io/bdnu2_v1

Kachergis, G., Marchman, V. A., & Frank, M. C. (2022). Toward a "Standard Model" of Early Language Learning. *Current Directions in Psychological Science*, *31*(1), 20–27. https://doi.org/10.1177/09637214211057836

Kachergis, G., Tan, A. W. M., Marchman, V. A., Dale, P. S., & Frank, M. C. (2025). *Measuring children's early vocabulary in low-resource languages using a Swadesh-style word list* (No. njm7d_v1). OSF.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2017). A Bootstrapping Model of Frequency and Context Effects in Word Learning. *Cognitive Science*, *41*(3), 590–622. https://doi.org/10.1111/cogs.12353

Kosie, J. E., & Lew-Williams, C. (2024). Infant-directed communication: Examining the many dimensions of everyday caregiver-infant interactions. *Developmental Science*, *27*(5), e13515. https://doi.org/10.1111/desc.13515

Long, B., Goodin, S., Kachergis, G., Marchman, V. A., Radwan, S. F., Sparks, R. Z., Xiang, V., Zhuang, C., Hsu, O., Newman, B., Yamins, D. L. K., & Frank, M. C. (2023). The BabyView camera: Designing a new head-mounted camera to capture children's early social and visual environments. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02206-1

Long, B., Sparks, R. Z., Xiang, V., Stojanov, S., Zi, Y., Keene, G., Tan, A. W. M., Feng, S. Y., Nag, A., Zhuang, C., Marchman, V. A., Yamins, D. L. K., & Frank, M. C. (2025, July). The BabyView dataset: High-resolution egocentric videos of infants' and young children's everyday experiences. *Proceedings of the 8th Annual Conference on Cognitive Computational Neuroscience*. https://doi.org/10.32470/vbbjtb0

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk.* (3rd ed.). Lawrence Erlbaum Associates.

Marchman, V. A., Dale, P., & Fenson, L. (2023). *MacArthur-Bates Communicative Development Inventories User's Guide and Technical Manual* (3rd edition). Brookes Publishing Co.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

Newman, R. S., Rowe, M. L., & Ratner, N. B. (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language*, *43*(5), 1158–1173. https://doi.org/10.1017/S0305000915000446

Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*(2). https://doi.org/10.1214/17-EJS1337SI

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision* (No. arXiv:2212.04356). arXiv. https://doi.org/10.48550/arXiv.2212.04356

Rowe, M. L. (2012). A Longitudinal Investigation of the Role of Quantity and Quality of Child-Directed Speech in Vocabulary Development. *Child Development*, *83*(5), 1762–1774. https://doi.org/10.1111/j.1467-8624.2012.01805.x

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663–12668. https://doi.org/10.1073/pnas.1419773112

Samuelson, L. K. (2021). Toward a Precision Science of Word Learning: Understanding Individual Vocabulary Pathways. *Child Development Perspectives*, *15*(2), 117–124. https://doi.org/10.1111/cdep.12408

Sepuri, T., Aw, K. L., Tan, A. W. M., Sparks, R. Z., Marchman, V. A., Frank, M. C., & Long, B. (2025, October). Characterizing young children's everyday activities using video question-answering models. *Proceedings of the Data on the Brain & Mind (DBM) Workshop at NeurIPS 2025*. https://doi.org/10.31234/osf.io/gndy9_v1

Sivula, T., Magnusson, M., & Vehtari, A. (2022). *Unbiased estimator for the variance of the leave-one-out cross-validation estimator for a Bayesian normal model with fixed variance* (No. arXiv:2008.10859). arXiv. https://doi.org/10.48550/arXiv.2008.10859

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568. https://doi.org/10.1016/j.cognition.2007.06.010

Suanda, S. H., Mugwanya, N., & Namy, L. L. (2014). Cross-situational statistical word learning in young children. *Journal of Experimental Child Psychology*, *126*, 395–411. https://doi.org/10.1016/j.jecp.2014.06.003

Tan, A. W. M., Loukatou, G., Braginsky, M., Mankewitz, J., & Frank, M. C. (2024). Predicting ages of acquisition for children's early vocabulary across 27 languages and dialects. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*.

Vong, W. K., Wang, W., Orhan, A. E., & Lake, B. M. (2024). Grounded language acquisition through the eyes and ears of a single child. *Science*, *383*(6682), 504–511. https://doi.org/10.1126/science.adi1374

Wang, A., Liu, L., Chen, H., Lin, Z., Han, J., & Ding, G. (2025). *YOLOE: Real-Time Seeing Anything* (No. arXiv:2503.07465). arXiv. https://doi.org/10.48550/arXiv.2503.07465

Wang, S., Chandra, A., Liu, A., Saligrama, V., & Gong, B. (2025). *BabyVLM: Data-Efficient Pretraining of VLMs Inspired by Infant Learning* (No. arXiv:2504.09426). arXiv. https://doi.org/10.48550/arXiv.2504.09426

Weizman, Z. O., & Snow, C. E. (2001). Lexical output as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology*, *37*(2), 265–279. https://doi.org/10.1037/0012-1649.37.2.265

Yang, J., Sepuri, T., Tan, A., Frank, M. C., & Long, B. (2025). Quantifying infants' everyday experiences with objects in a large corpus of egocentric videos. *Extended Abstracts of the 8th Annual Conference on Cognitive Computational Neuroscience*.

Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., Jin, P., Zhang, W., Wang, F., Bing, L., & Zhao, D. (2025). *VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding* (No. arXiv:2501.13106). arXiv. https://doi.org/10.48550/arXiv.2501.13106