Does Contextual Informativeness Predict Preschoolers' Word Learning from Stories?

Maria Valentini[1,4], **Julisa Granados**[2], **Haley McVay**[2], **Katharina von der Wense**[1,3,4] **& Eliana Colunga**[1,2,4]
[1]Department of Computer Science, University of Colorado Boulder
[2]Department of Psychology and Neuroscience, University of Colorado Boulder
[3]Department of Computer Science, Johannes Gutenberg University Mainz
[4]Institute of Cognitive Science, University of Colorado Boulder
{first.last}@colorado.edu

Abstract

There is a strong relationship between book-rich environments and vocabulary size in early childhood, and for preschoolers, storybook sharing remains an important source of new vocabulary. In this work we ask, what makes a story an effective tool for word learning and vocabulary enrichment? We use data from 49 preschoolers sharing AI-generated picture books with a caregiver, with each set of stories containing 20 target words (nouns, verbs, and adjectives), to assess the feasibility of using both linguistic and visual contextual informativeness metrics to evaluate the quality of stories as support for learning new vocabulary. Results show that contextual informativeness impacts learning differently across word types: visual metrics and linguistic ground truth measures both correlate with learning for nouns and verbs (but not adjectives), and our automated metric for approximating linguistic informativeness shows significant predictive performance specifically for verbs. These results speak to the importance of considering different sources of information — including linguistic and visual information — when designing materials that support learning. We discuss these findings in the context of improving automatically generated child-directed stories to support the learning of a variety of different types of words.

Keywords: Contextual informativeness; shared story reading; word learning; early childhood; large language models

Background

Research has shown that gaps in vocabulary size between children begin to develop as early as 18 months (Brushe et al., 2021) and continue through early childhood and beyond. For example, Biemiller and Slonim (2001) estimate the vocabulary of 4th graders in the lowest quartile to be less than half the size that of 4th graders in the highest quartile. These early differences are especially important because, alongside oral language, phonological awareness, fluency, and comprehension, vocabulary is one of several interdependent skills needed to support children's successful reading development (Vellutino et al., 2007). While these skills develop interdependently, research suggests that vocabulary plays a particularly important role in supporting oral language and, consequently, reading comprehension (Anderson & Freebody, 1979).

In turn, reading comprehension is a critical predictor of children's success across many areas of academic achievement, including science, mathematics, and the social sciences (Bastug, 2014). In line with this, differences in vocabulary size are often reflected in differences in other areas of language and academic achievement: those children who fall behind prior to school entry continue to demonstrate lower reading, writing, and oral language skills through middle and high school than their peers (Rescorla, 2009). Given the apparent downstream impacts of vocabulary size, experts suggest that intervening early and increasing access to vocabulary-enriching materials is critical (de Bondt et al., 2020).

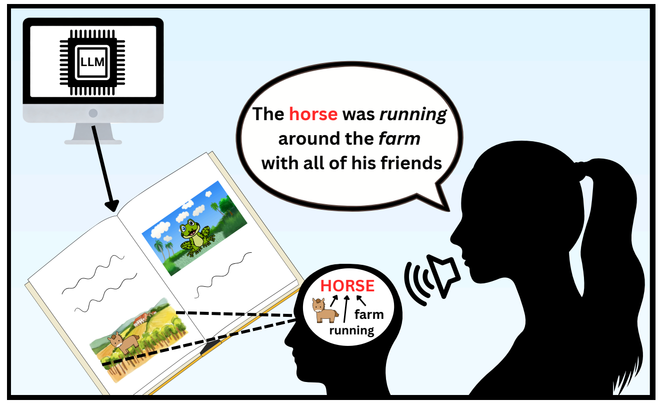

Figure 1: A child simultaneously receiving visual and linguistic information relating to the word 'horse' from an LLM-generated storybook.

Shared reading—reading a storybook with an adult caregiver—has been shown to be effective in producing vocabulary gains in young children (Mol et al., 2008, 2009). Shared reading supports vocabulary development by providing children with opportunities to engage in interactive dialogue with caregivers. This dialogue is often supported by caregivers' use of dialogic reading behaviors, such as asking questions to encourage child participation and providing responsive feedback to scaffold children's responses (Petrie et al., 2023).

Shared reading may also support vocabulary development through the linguistically rich input provided by the books themselves. For example, storybooks often contain more sophisticated vocabulary than children typically encounter in their everyday environments (Montag et al., 2015). Together, these qualities make shared reading a particularly effective setting for early vocabulary learning. Consequently, much intervention work has focused on introducing new words through story reading (Flack et al., 2018; Marulis & Neuman, 2010).

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

Automatic evaluation of stories Until recently, enabling educators and caregivers to implement individualized vocabulary enrichment at scale seemed unreasonable, particularly given the manual labor required to design individual language materials for each child. Recent work, however, investigates the ability of large language models (LLMs) to address this problem computationally through the generation of stories, which can be augmented to contain desirable target words (Peng et al., 2023; Valentini et al., 2023; Weber et al., 2024). In addition to targeted language learning materials, LLM generation allows for personalization to fit children's interests and make connections with their real lives, which has been shown to improve word learning (Zheng et al., 2022). These works also reveal several major weaknesses, emphasizing the need for effective automatic evaluation of generated materials before these methods can be widely adopted. Various works have looked at the evaluation of aspects of LLM-generated stories such as coherence, relevance, and engagement (Chhun et al., 2022), characteristics important for stories that are effective in supporting children's word learning. Valentini et al. (2023) assessed the simplicity of generated stories, finding that generation models sometimes use overly complicated language even when specifically prompted, and solved this issue by including an external simplification model in the pipeline.

Contextual informativeness and word learning Another component of text generated for vocabulary enrichment which has been evaluated in recent work is the contextual informativeness of the text surrounding target words. Frishkoff et al. (2011) demonstrated that context is important for adults learning word meanings. They augmented contexts by hand and repeatedly asked adults to define the meaning of included target words, showing additional contexts over time and measuring the semantic closeness of each new definition, finding that better context led to better definitions.

The influence of contextual informativeness on word learning in children has received less attention. However, work by Hadley et al. (2015) suggests that preschool-aged children may develop deeper word knowledge when novel vocabulary is taught with rich semantic support. In their study, preschoolers participated in a shared reading and play-based intervention in which novel words were taught with rich semantic information (e.g., child-friendly definitions, picture cues, and other meaningful contexts). Children showed significant growth in vocabulary depth from pre- to post-intervention. Similarly, Weber and Colunga (2022) compared toddler's word learning from picture books depicting target words in rich semantic contexts versus as isolated pictures on a white background, and found that visual context made no difference for typically developing children, but late talkers learned more in the rich visual context condition. These results suggest that both linguistic and visual informativeness in picture books may influence children's vocabulary learning (see Figure 1).

Given the demonstrated importance of context for incidental word learning, a recent body of work examines automatic evaluation methods for contextual informativeness. Kapelner et al. (2018) measured the contextual informativeness of adult data by having annotators rate each context on a Likert scale and develop a binary classifier to automatically classify rich vs. poor contexts, finding that teachers prefer the pre-filtered context 60.7% of the time. Valentini et al. (2025) specifically focused on the task of automatically evaluating contextual informativeness in stories generated for early childhood vocabulary support, developing a metric and demonstrating results alongside a semantic-space-narrowing human annotation task.

The current study Weber et al. (2024) conducted a study evaluating the educational potential of LLM-generated stories, running shared reading experiments to determine whether or not children are able to learn new words from stories written by LLMs. The original LLM-generated stories were developed in Valentini et al. (2023) and consist of 240 stories with 5 target words each. Prompts were designed to elicit short and simple stories for teaching words to preschool-aged children, and words were distributed across nouns, verbs, and adjectives. Weber et al. selected a 32-story subset from the original 240 based on appropriateness, learnability, and coherence. They then illustrated the stories using a variety of methods with and without AI generation and printed them into spiral-bound, laminated storybooks. They collected data from 49 preschooler-caregiver pairs, each given a set of 4 AI-generated picture books for shared reading. Each book contained 5 words for a total of 20 target words, including nouns, verbs, and adjectives. In addition to finding that parents considered automatically generated stories engaging, age-appropriate, and educational, they found that children performed significantly better when asked to name pictures depicting studied target words compared to not studied control words of similar difficulty that served as target words for other participants. They also observed an effect of word type on children's learning, with nouns being more readily learned than verbs and adjectives. Overall, their work highlights the potential of using LLMs to generate stories based on specific target words as a method for enriching early vocabulary.

In this paper, we utilize data from Weber et al.'s study and perform a series of statistical analyses on the impact of both linguistic (informativeness of surrounding text) and visual (image depictions of the target word) cues in the picture books to examine the relationships between these informational factors and word learning outcomes. We further probe these interrelations using a generalized linear mixed-effects model with the fixed effects of word type (noun/verb/adjective) to assess the impact of these cues for different types of words. These analyses allow us to probe relationships between the information included in automatically generated storybooks and their effectiveness in teaching new words to preschoolers.

Methods

Data

Data for these experiments, including child word learning data as well as the corresponding informativeness scores of the linguistic and visual contexts for each target word, was collected from several primary sources.

Child Learning Data The learning data here comes from Weber et al. (2024). As described above, researchers gave out 4 books each to 49 preschool-aged children (ages 3-6) and their caregivers, resulting in each child being assigned a set of 20 target words, including nouns, verbs, and adjectives. To measure learning, children were asked to identify images depicting each of their 20 target words. The data we use from this study contains learning results from 157 story/target-word pairings, with each target word appearing in 1 or 2 storybooks. The learning data was used in conjunction with various measures of the contextual informativeness offered for each target word in its corresponding picture book, as described below, to carry out the analyses presented here.

Context Annotations Linguistic context annotations were taken from Valentini et al. (2025), who turn each generated story into Cloze task (Taylor, 1953)-style prompts in order to obtain human guesses and approximate each story's contextual informativeness with respect to contained target words (see Figure 2 for an example). The authors blocked out all 5 target words from each prompt in order to simulate for annotators the viewpoint of a child unfamiliar with any of the target words. Gold standard contextual informativeness scores were calculated by averaging the cosine similarities between the annotators' guesses and the actual missing target words, with the goal of measuring how well the context allows readers to narrow down the semantic space.

Visual context was measured by determining how many times each of the target words was depicted in each picture book. Target words were depicted between 0 and 10 times per picture book, with an average of 4.38 depictions per target word overall (McVay, 2024).

Other Measures Finally, information on the difficulty of words is obtained from Kuperman et al. (2012), a dataset containing the average reported age of acquisition (AoA) for 35,000 English words. We combined the stories with the child word-learning data and adult human guesses, at which point we determined scores using the automated Gemini-based metric described in Valentini et al. (2025). We prompted Gemini-2.0-flash with the Cloze-task version of each story and calculated the cosine similarity between its guess and the missing target word, aggregating over 5 responses for each sample. This resulted in a dataset with both contextual metrics and full child learning data in order to allow direct comparison. It contains 157 target-word/story pairings. Of these 157 words, 45 appear in two stories, meaning there are 112 unique target words: 42 nouns, 34 verbs, and 36 adjectives.

Analyses

All analyses were conducted in R. They include correlation measures, partial correlation measures, mixed-effects model results, and estimated marginal trends.

Once upon a time, there was a ______ man named Joe who lived in a ______ house. He loved to paint, and every day he would go outside to his ______ and begin ______. He would always tie a <mask> around his head to keep the sun off his ______ head. Joe was known for his beautiful paintings of the ______ sky and the colorful flowers that surrounded his house.

Figure 2: Short example of a cloze-style prompt in which no target words are known, and participants are asked to guess the word replaced by the <mask> token.

Correlations In order to address our question of how target word contextual informativeness relates to child learning of that target word, we first took correlations between the correct answer proportion for each word/story and that sample's human-provided and automatic contextual informativeness scores. This provided a straightforward method for comparison, but it lacked some nuance in accounting for other major factors such as individual target-word difficulty.

To solve this, we additionally conducted partial correlations where we controlled for each target word's age of acquisition. These correlations were collected both overall (all words combined) and separately by word type, as some methods may be better at predicting child performance for some word types more than others. This provided a more complete picture of interactions and allowed us to isolate specific effects for further exploration, but it still treated the proportion of correctly produced words as an independent continuous variable and did not account for item-level dependence or differences in measurement precision.

Mixed-Effects Models We further analyzed word learning outcomes using a *generalized linear mixed-effects model* with a binomial error distribution and logit link, solving the non-independence issues brought on by partial correlations. The dependent variable was the number of correct responses for each target–story pairing, modeled as a two-column response cbind(k_{correct}, n_{total} - k_{correct}) where k_{correct} denotes the number of correct responses and n_{total} the total number of responses for that item.

Fixed effects included *word type* (noun [reference], verb, adjective), *metric* (Gemini similarity [reference] vs. human similarity), and *value*, the similarity score associated with the given measure (z-scored). All two-way and three-way interactions among word type, metric, and value were included to allow the relationship between similarity and accuracy to vary by word type and by similarity metric.

To control for item-level difficulty, we additionally included a binomial mixed-effects model with target age of acquisition (target AoA) as a main-effect covariate. A formal definition for the general binomial model k_{ij} is as follows:

$$\begin{aligned} k_{ij} &\sim \text{Binomial}(n_{ij}, p_{ij}), \\ \text{logit}(p_{ij}) &= X_{ij}\boldsymbol{\beta} + u_j, \quad u_j \sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{1}$$

where, for our use case, X_{ij} contains word type, metric, value, and their interactions.

This formulation models the probability of a correct response as a function of the fixed predictors, while accounting for item-level dependence induced by repeated observations of the same target–story pairing with the random intercept term u_j. For each observation i within the target-story grouping j, the number of correct responses k_{ij} is assumed to follow a binomial distribution with parameters n_{ij} (the total number of responses) and p_{ij} (the probability of a correct response). The probability p_{ij} is linked to a linear predictor via the logit link function. The linear predictor, given by $X_{ij}\boldsymbol{\beta} + u_j$, was specified in R as cbind(k_correct, n_total - k_correct) ~ word type * metric * value + (1 | TargetStory) in the initial model, and cbind(k_correct, n_total - k_correct) ~ word type * metric * value + target AoA + (1 | TargetStory) in the augmented.

Estimated Marginal Trends Because our primary question concerns how predictive human-produced and automatically generated contextual informativeness scores are within each word type × metric setting, we additionally examined estimated marginal trends (EMTs) of the z-scored informativeness value. These trends quantify the conditional slope of informativeness value on accuracy (on the log-odds scale), separately for each combination of lexical category and similarity measure, averaging over the remaining predictors in the model. In total, we report six distinct estimates of how informativeness relates to correctness: one slope per word type × metric. Each reported slope reflects the change in log-odds of correctness per standard deviation of informativeness.

Table 1: Full and partial correlations between informativeness metrics and target accuracy. Full correlations show the raw relationship, while partial correlations control for target word age of acquisition. Correlations are computed across Target/Story items within each lexical category.

Full Correlations

Word type	Metric	r	t	p	n
Overall	Human	0.220	2.81	0.006	157
Overall	Gemini	0.153	1.93	0.055	157
Noun	Human	0.230	1.81	0.075	61
Noun	Gemini	−0.021	−0.16	0.870	61
Verb	Human	0.249	1.74	0.088	48
Verb	Gemini	0.432	3.25	0.002	48
Adjective	Human	0.083	0.56	0.577	48
Adjective	Gemini	−0.029	−0.20	0.846	48

Partial Correlations (Controlling for AoA)

Word type	Metric	r	t	p	n
Overall	Human	0.214	2.72	0.007	157
Overall	Gemini	0.116	1.45	0.148	157
Noun	Human	0.225	1.76	0.084	61
Noun	Gemini	−0.036	−0.27	0.786	61
Verb	Human	0.245	1.70	0.097	48
Verb	Gemini	0.380	2.76	0.008	48
Adjective	Human	−0.008	−0.06	0.956	48
Adjective	Gemini	−0.049	−0.33	0.746	48

Results

Correlations

Linguistic Context For a complete list of linguistic correlation results, see Table 1. Both full and partial correlation results demonstrated relationships between contextual informativeness metrics and child performance, and that these relationships vary a great deal by factors such as word type.

Looking first at the human-scored gold standard data to observe the general effects of contextual informativeness, we found a significant overall effect across words. The full and partial correlations recorded p-values of approximately 0.0056 and 0.0073, respectively, indicating a relationship between how informative the context surrounding a target word is and how well children perform in identification tasks. Split by word type, the p-value dropped due to the relatively smaller number of samples, but the differences in correlation allowed us to observe trends across the types. The base partial correlation value of $r_{cp \cdot a} = 0.214$ (with c = context score, p = proportion correct, and a = target AoA) increased for verbs and nouns to $r_{cp \cdot a} \approx 0.245$ and $r_{cp \cdot a} \approx 0.225$, respectively. Adjectives showed no correlation or significance, with an $r_{cp \cdot a}$ value of −0.008.

Regarding our goal of finding a metric to automatically predict how well a child may learn a word from a story, the Gemini scoring method was found to have a significant relationship with correctness scores for verbs, with $p \approx 0.008$ in spite of the low number of samples, and a partial correlation value of $r_{cp \cdot a} = 0.380$. Though this relationship does not hold for nouns or adjectives, it suggests that there may be measurable factors to predict and optimize the learning potential for at least some types of target words.

Visual Context Illustration results are included in full in Table 2. They demonstrated that visual context for vocabulary words, while not as predictive as linguistic information, still has a significant relationship with child word learning performance. The partial correlation using target word AoA as a covariate had slightly more significance than the full correlation, unlike in linguistic context results ($p \approx 0.039$ vs. $p \approx 0.041$). This suggests that differences in target-word difficulty may have more of an impact on how children learn from context clues than how they learn from illustrations.

Splitting the visual analysis by word type showed that visual context is more important for verbs and nouns than it is for adjectives: the base partial correlation of $r_{cp \cdot a} \approx 0.165$ (with c = context score, p = proportion correct, and a = target AoA) increases for nouns and verbs to 0.196 and 0.188, respectively. As with linguistic context, visual context demonstrated no relationship with child word learning for adjectives, with $r_{cp \cdot a} \approx 0.028$.

Table 2: Correlations between illustrations and accuracy.

Full Correlations

Word type	r	t	p	n
Overall	0.163	2.06	0.041	157
Noun	0.204	1.60	0.114	61
Verb	0.271	1.91	0.063	48
Adjective	−0.012	−0.08	0.934	48

Partial Correlations (Controlling for AoA)

Word type	r	t	p	n
Overall	0.165	2.08	0.039	157
Noun	0.196	1.52	0.133	61
Verb	0.188	1.28	0.205	48
Adjective	0.028	0.19	0.852	48

Mixed-Effects Models

Full results of the initial binomial mixed-effects model can be found in Table 3. They demonstrated that none of the fixed effects of interest were statistically reliable. This pattern suggested that a large proportion of variance in accuracy was more driven by item-level difficulty than it was by systematic differences associated with similarity metrics or category, so controlling for item-level difficulty effects would be necessary to examine additional effects of context.

After controlling for target word AoA, however, a main effect of word type became evident: adjectives and verbs are both associated with considerably lower accuracy than nouns. Full results from this model can be found in Table 4.

The estimate values of $\beta \approx -0.621$ for verbs and $\beta \approx -0.727$ for adjectives (with $p \approx 0.085$ and $p \approx 0.041$) represent the difference in log-odds of a correct response when switching the target's word type from the reference (noun). Calculating the value of e^{β} yields the odds ratio, which has values of approximately 0.537 and 0.483 for verbs and adjectives, respectively. This means that by switching the target from a noun to a verb, the odds of a correct answer drop by around 46%. Switching from a noun to an adjective means the odds drop nearly 52%.

Estimated Marginal Trends

Results from estimated marginal trends, shown in full in Table 5, exhibit directional consistency and further support the claims we derive from correlation results.

The trend column in the table represents the estimated slope of the z-scored values within each setting, where the unit is change in log-odds of correctness per 1 standard deviation increase in similarity. For example, looking at verbs, a 1 SD increase in Gemini similarity changes the log-odds of correctness by +0.2645. This yields an odds ratio of 1.303, so a child is approximately 30% more likely to get a verb correct if there is 1 SD more context (according to the Gemini metric). We note, however, that these results are not statistically significant on their own and would require additional testing to confirm. This is not unusual: because the random intercept absorbs much of the between-item variance that drives the partial correlations, the mixed model represents a more conservative test and is expected to yield weaker fixed-effect estimates under conditions of noisy child data.

Table 3: Mixed-effects logistic regression predicting target word accuracy from similarity metric type (Gemini [reference] vs. Human), standardized similarity value, and word type (Noun [reference], Verb, Adjective). Random intercepts were included for each target story item. *Sim* refers to similarity value.

Base Model

Fixed Effect	β	SE	z	p
(Intercept)	−0.20	0.24	−0.81	.416
Word type: Verb	−0.49	0.38	−1.32	.188
Word type: Adjective	−0.50	0.37	−1.36	.173
Metric: Human	−0.01	0.18	−0.06	.955
Similarity Value (z)	0.02	0.18	0.11	.915
Verb × Human	0.04	0.29	0.14	.892
Adjective × Human	0.04	0.31	0.12	.906
Verb × Sim	0.26	0.25	1.02	.308
Adjective × Sim	−0.02	0.26	−0.08	.937
Human × Sim	0.11	0.23	0.50	.617
Verb × Human × Sim	−0.18	0.32	−0.58	.562
Adjective × Human × Sim	−0.05	0.40	−0.13	.895
Random Effects				
Target/Story (Intercept): σ^2 = 2.50, SD = 1.58				
N_{obs} = 314, N_{groups} = 157				

Even without statistical significance, however, these results support the results of our full and partial correlations, demonstrating directional consistency across settings. Examining the coefficients, there are three main findings: human similarity slopes are consistently positive for nouns and verbs, Gemini similarity is positive for verbs but null elsewhere, and adjectives are null regardless of similarity metric.

Discussion

Although our results vary in their individual significance, they follow patterns that remain internally consistent as well as consistent with existing psycholinguistic theories.

Our first major result was that linguistic contextual informativeness does appear to be helpful for learning words overall. Full and partial correlations both showed that a significant relationship exists between the human gold standard for contextual informativeness and child word learning performance (p < 0.008). Breaking the results down by word type, we found more specifically that contextual information helps with verbs and nouns but does not demonstrate an effect for adjectives. This pattern was also reflected in model results, where EMT coefficients were all directionally consistent.

We also found that children have a more difficult time learning verbs and adjectives in general than they do nouns. This

Table 4: Mixed-effects logistic regression predicting target word accuracy from similarity measure type (Gemini [reference] vs. Human), standardized similarity value, and lexical category (Noun [reference], Verb, Adjective), controlling for target word age of acquisition. Random intercepts were included for each target story item.

AoA-Controlled Model

Fixed Effect	β	SE	z	p
(Intercept)	−0.10	0.23	−0.44	.661
Word type: Verb	−0.62	0.36	−1.72	.085 ·
Word type: Adjective	−0.73	0.36	−2.04	.041*
Metric: Human	−0.01	0.18	−0.07	.941
Similarity Value (z)	−0.01	0.18	−0.07	.943
Target AoA (z)	−0.63	0.15	−4.31	< .001***
Verb × Human	0.04	0.29	0.14	.887
Adjective × Human	0.00	0.31	0.02	.987
Verb × Sim	0.28	0.25	1.11	.269
Adjective × Sim	−0.01	0.26	−0.04	.966
Human × Sim	0.13	0.22	0.57	.568
Verb × Human × Sim	−0.18	0.32	−0.56	.573
Adjective × Human × Sim	−0.13	0.40	−0.32	.752
Random Effects				
Target/Story (Intercept): $\sigma^2 = 2.18$, SD = 1.48				
$N_{obs} = 314$, $N_{groups} = 157$				

$\cdot\ p < .1, ^{*}p < .05, ^{***}p < .001$

Table 5: Estimated marginal trends for the effect of similarity value on target accuracy by lexical category and measure type. Trends represent the estimated change in log-odds of accuracy per standard deviation increase in similarity value. Models without and with age of acquisition (AoA) control are shown.

Base Model

Word type	Metric	Trend	SE	95% CI
Noun	Gemini	0.020	0.184	[−0.341, 0.380]
Noun	Human	0.133	0.221	[−0.300, 0.565]
Verb	Gemini	0.280	0.177	[−0.067, 0.626]
Verb	Human	0.208	0.189	[−0.162, 0.578]
Adjective	Gemini	−0.001	0.186	[−0.366, 0.364]
Adjective	Human	0.059	0.348	[−0.624, 0.742]

AoA-Controlled Model

Word type	Metric	Trend	SE	95% CI
Noun	Gemini	−0.013	0.180	[−0.365, 0.340]
Noun	Human	0.115	0.216	[−0.309, 0.539]
Verb	Gemini	0.265	0.175	[−0.079, 0.608]
Verb	Human	0.214	0.187	[−0.153, 0.581]
Adjective	Gemini	−0.024	0.184	[−0.385, 0.337]
Adjective	Human	−0.023	0.346	[−0.700, 0.655]

was evident from our adjusted mixed model results, in which switching a target word from a noun to a verb or adjective (holding all other factors constant) drops the probability of a child guessing the word correctly by around 50%. This is in line with the fact that, in early vocabularies, nouns are more prevalent than verbs and adjectives, suggesting that they are easier to learn (McDonough et al., 2011).

Results from our visual information analyses suggested that the overall effect of illustrations is significant ($p < 0.04$). The word type-based analysis showed more specifically that illustrations appear to help with learning nouns and verbs, but do not demonstrate an effect for adjectives. We theorize this is possibly because of the difficulty of illustrating an adjective, but this remains an open question.

In searching for an automatic metric to assess the educational quality of generated context with respect to target words, we found that the Gemini scoring method is a reasonable predictor of verb learning. Controlling for other variables like target AoA makes the relationship slightly weaker, but it remains considerable and significant. As with human metric results, we observed that the same patterns in the correlations extended to the mixed-effects model, albeit with less statistical power. This relationship, however, does not extend beyond the verb word type. It may be weak for other word types partially due to the nature of the production task: nouns and verbs are clearly depicted in images and were probed with the simple questions, "what is this?" or "what is s/he doing?" respectively, but adjective trials require contrast scaffolding: "this one is smooth, this one is?". Verbs, being generally harder to learn than nouns, benefit from the inclusion of richer context.

Our final primary finding across analyses was that adjectives showed essentially no relationship with context, linguistic or visual. In addition to suggesting that surrounding context is relatively unimportant for learning adjectives, this result brings about more questions regarding the nature of learning adjectives in general; it seems likely that adjectives are just very hard to teach effectively using these methods, and they may require a different type of contextual support.

The findings of this study have several implications: for one, they can inform the development of automatic story generation pipelines, placing emphasis on including informative contexts for nouns and verbs. Additionally, while the results here are demonstrated on LLM-generated stories, the analysis is conducted irrespective of author (based solely on embedded textual features), and its conclusions thus likely generalize to human-authored texts and provide insight about the broader nature of incidental word learning in children.

In all, the results of the present study demonstrate that the effects of linguistic and visual contextual information on word learning vary considerably by factors like word type, but significant trends nonetheless emerge and are reflected across analyses. Both human-annotated and automatically generated metrics show promise for evaluating verb-learning contexts, but the null findings for adjectives suggest that some word types may require fundamentally different pedagogical approaches. In addition to emphasizing the need to consider all modalities of input when evaluating stories for their word learning potential, these results highlight some of the ways in which children learn different types of words differently and provide additional support for existing linguistic theories.

References

Anderson, R. C., & Freebody, P. (1979, August). *Vocabulary knowledge* (Technical Report No. 136). Center for the Study of Reading, University of Illinois at Urbana-Champaign. https://eric.ed.gov/?id=ED177480

Bastug, M. (2014). The structural relationship of reading attitude, reading comprehension and academic achievement. *International Journal of Social Science & Education*, *4*(4), 931–946. https://ijsse.com/sites/default/files/issues/2014/v4-i4-2014-1/Paper-20.pdf

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, *93*, 498–520. https://doi.org/10.1037/0022-0663.93.3.498

Brushe, M. E., Lynch, J., Reilly, S., Melhuish, E., Mittinty, M. N., & Brinkman, S. A. (2021). The education word gap emerges by 18-months: Findings from an australian prospective study. *BMC pediatrics*, *21*(1), 247. https://doi.org/10.1186/s12887-021-02712-1

Chhun, C., Colombo, P., Suchanek, F. M., & Clavel, C. (2022, October). Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, & S.-H. Na (Eds.), *Proceedings of the 29th international conference on computational linguistics* (pp. 5794–5836). International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.509/

de Bondt, M., Willenberg, I. A., & Bus, A. G. (2020). Do book giveaway programs promote the home literacy environment and children's literacy-related behavior and skills? *Review of Educational Research*, *90*(3), 349–375. https://doi.org/10.3102/0034654320922140

Flack, Z. M., Field, A. P., & Horst, J. S. (2018). The effects of shared storybook reading on word learning: A meta-analysis. *Developmental Psychology*, *54*(7), 1334–1346. https://doi.org/10.1037/dev0000512

Frishkoff, G. A., Perfetti, C. A., & Collins-Thompson, K. (2011). Predicting robust vocabulary growth from measures of incremental learning. *Scientific Studies of Reading*, *15*(1), 71–91. https://doi.org/10.1080/10888438.2011.539076

Hadley, E. B., Dickinson, D. K., Hirsh-Pasek, K., Golinkoff, R. M., & Nesbitt, K. T. (2015). Examining the acquisition of vocabulary knowledge depth among preschool students. *Reading Research Quarterly*, *51*(2), 181–198. https://doi.org/10.1002/rrq.130

Kapelner, A., Soterwood, J., Nessaiver, S., & Adlof, S. (2018). Predicting contextual informativeness for vocabulary learning. *IEEE Transactions on Learning Technologies*, *11*(1), 13–26. https://doi.org/10.1109/TLT.2018.2789900

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children's word learning: A meta-analysis. *Review of Educational Research*, *80*(3), 300–335. https://doi.org/10.3102/0034654310377087

McDonough, C., Song, L., Pasek, K. H., Golinkoff, R. M., & Lannon, R. (2011). An image is worth a thousand words: Why nouns tend to dominate verbs in early word learning. *Developmental science*, *14*(2), 181–189. https://doi.org/10.1111/j.1467-7687.2010.00968.x

McVay, H. (2024). An evaluation of ai-generated children's stories: Do word characteristics correlate with word learning?

Mol, S. E., Bus, A. G., & de Jong, M. T. (2009). Interactive book reading in early education: A tool to stimulate print knowledge as well as oral language. *Review of Educational Research*, *79*(2), 979–1007. https://doi.org/10.3102/0034654309332561

Mol, S. E., Bus, A. G., de Jong, M. T., & Smeets, D. J. H. (2008). Added value of dialogic parent–child book readings: A meta-analysis. *Early Education and Development*, *19*(1), 7–26. https://doi.org/10.1080/10409280701838603

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science*, *26*(9), 1489–1496. https://doi.org/10.1177/0956797615594361

Peng, Z., Wang, X., Han, Q., Zhu, J., Ma, X., & Qu, H. (2023). Storyfier: Exploring vocabulary learning support with text generation models, 1–16. https://doi.org/10.1145/3586183.3606786

Petrie, A., Mayr, R., Zhao, F., & Montanari, S. (2023). Parent-child interaction during storybook reading: Wordless narrative books versus books with text. *Journal of Child Language*, *50*(1), 104–131. https://doi.org/10.1017/S0305000921000763

Rescorla, L. (2009). Age 17 language and reading outcomes in late-talking toddlers: Support for a dimensional perspective on language delay. *Journal of speech, language, and hearing research: JSLHR*, *52*(1), 16–30. https://doi.org/10.1044/1092-4388(2008/07-0171)

Taylor, W. L. (1953). "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, *30*(4), 415–433.

Valentini, M., Weber, J., Salcido, J., Wright, T., Colunga, E., & von der Wense, K. (2023, December). On the automatic generation and simplification of children's stories. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 3588–3598). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.218

Valentini, M., Wright, T., Marashian, A., Weber, J., Colunga, E., & von der Wense, K. (2025, January). Measuring contextual informativeness in child-directed text. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio,

& S. Schockaert (Eds.), *Proceedings of the 31st international conference on computational linguistics* (pp. 8109–8120). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.540/

Vellutino, F. R., Tunmer, W. E., Jaccard, J. J., & Chen, R. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific Studies of Reading*, *11*(1), 3–32. https://doi.org/10.1080/10888430709336632

Weber, J., & Colunga, E. (2022). Learning from word books: Does the type of illustration matter? *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). https://escholarship.org/uc/item/5kv5g42x

Weber, J., Valentini, M., Wright, T., von der Wense, K., & Colunga, E. (2024). Evaluating llms as tools to support early vocabulary learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*(0). https://escholarship.org/uc/item/7v69f0dj

Zheng, L., Long, M., Zhong, L., & Gyasi, J. F. (2022). The effectiveness of technology-facilitated personalized learning on learning achievements and learning perceptions: A meta-analysis [ERIC Number: EJ1352870]. *Education and Information Technologies*, *27*(8), 11807–11830. https://doi.org/10.1007/s10639-022-11092-7