

Pathways to Early Vocabulary Acquisition*

Mapping English Language Development Trajectories in Children Aged 16–30 Months

Yongqi Liu

November 28, 2024

This study examines vocabulary acquisition in children aged 16–30 months using data from the MacArthur-Bates Communicative Development Inventories. Age is the strongest predictor, with vocabulary size increasing consistently and foundational categories like Function Words, Living Things, and Objects showing higher predicted probabilities, indicating their universal acquisition. Complex categories such as Sensory Words, Adjectives, and Places exhibit broader distributions and lower probabilities, reflecting their reliance on contextual understanding and cognitive complexity. These findings underscore age-specific patterns in vocabulary growth and the need for targeted interventions to support diverse language learning trajectories.

Table of contents

1	Introduction	2
2	Data	4
2.1	Overview	4
2.2	Measurement	5
2.3	Outcome Variable	6
2.3.1	High Vocabulary Score	6
2.4	Predictor Variables	8
2.4.1	Age	8
2.4.2	Word Category	8
2.4.3	Norming Status	11

*Code and data are available at: https://github.com/Cassieliu77/Vocabulary_Learning_Pattern.git

3	Model	11
3.1	Model Overview	11
3.2	Model Assumptions	12
3.3	Interpretation of Coefficients	13
3.4	Model Justification	14
4	Results	14
4.1	Variability in Production Vocabulary	14
4.2	Median Vocabulary Size Change by Age	15
4.3	Prediction for the Probability of High Vocabulary Level	16
4.4	Predicted Probabilities by Age	16
4.5	Distribution of Predicted Probabilities by Word Category	18
5	Discussion	19
5.1	Age and Developmental Dynamics	19
5.2	Word Categories: Patterns and Variability	19
5.3	Implications for Early Childhood Education	21
5.4	Limitations and Future Directions	22
A	Appendix	23
A.1	Survey Design: The CDI Framework	23
A.2	Sampling Framework	24
A.3	Observational Nature of the Data	24
A.4	Connecting to the Literature	24
A.5	Simulation and Recommendations	25
B	Model details	25
B.1	Model Summary	25
B.2	Diagnostics	25
B.2.1	Confusion Matrix	25
B.2.2	ROC Curve and AUC	27
B.2.3	Precision-Recall Curve	27
B.2.4	Binned Residual Plot	28
C	Acknowledgements	28
	References	29

1 Introduction

Language development in infants and toddlers is often discussed in terms of isolated stages and distinct milestones. However, the fact is that language grows continuously, progressing month

by month—and even day by day—in between these milestones. While language milestones are important, it is also crucial to recognize the month-by-month progression in comprehension and production during early language development. To better understand this growth, this study shows investigation on CDI (2024) data and illustrates language development monthly over 16-30 months aged children.

Understanding how children acquire vocabulary is a cornerstone to develop linguistics and early childhood education. Vocabulary growth is not only a key indicator of cognitive development but also serves as a foundation for future linguistic and academic success. The interplay of factors such as age, linguistic environments, and lexical categories influences children’s vocabulary acquisition in complex and dynamic ways. By analyzing patterns and predictors of vocabulary growth, this paper describes the developmental trajectories that underpin language acquisition, offering a clearer understanding of how children progress in their linguistic abilities.

This study utilizes a dataset from the Braginsky (2024) package, which aggregates data from CDI (2024), to investigate vocabulary development in early childhood. By employing a logistic regression model, the research also estimates the likelihood of achieving a high vocabulary score using predictors such as age, norming status, and lexical categories. This model serves as a tool to explore how these factors collectively influence children’s vocabulary acquisition trajectories.

The estimand of interest in this study is the probability that an English-speaking child achieves a high vocabulary score, defined by surpassing a threshold of 300 combined comprehension and production words. These predictors include the child’s age (measured in months), their inclusion in a norming group (standardized benchmark), and the lexical category of the words they acquire (e.g., Function Words, Adjectives, Verbs). The study aims to quantify the influence of these factors on vocabulary proficiency and identify how age and lexical categories uniquely shape developmental trajectories. The model further investigates variability in vocabulary acquisition across categories, highlighting disparities between foundational categories, such as Function Words, which are more consistently acquired, and complex categories like Sensory Words, which show broader variability due to contextual and cognitive demands.

Key findings reveal distinct trajectories of vocabulary acquisition across different word categories, with age emerging as the strongest predictor of vocabulary growth. Also, the acquisition speed among children differs a lot. The analysis shows that as children age, their likelihood of achieving high vocabulary scores increases consistently, reflecting the natural progression of language development. Foundational categories such as “Living Things” and “Function Words” exhibit stable and high predicted probabilities, likely due to their frequent use in daily communication and their role in forming early linguistic structures. In contrast, complex categories such as “Sensory Words” and “Adjectives” display greater variability, highlighting the cognitive and contextual challenges associated with their acquisition. Additionally, the influence of linguistic environments on vocabulary development is significant. Children in the norming groups, who benefit from structured linguistic exposure and are assessed under more controlled and standardized conditions, consistently outperform their non-norming peers in

vocabulary scores. This highlights the impact of structured environments on language development and underscores the importance of norming as a tool for calibrating developmental assessments. The results suggests the need for targeted interventions, particularly in supporting the acquisition of underdeveloped categories such as Sensory Words and Adjectives. Furthermore, the findings in this paper pave the way for future research to explore broader contextual factors—such as socio-economic status and bilingualism—that may further explain the observed variability in children’s language acquisition trajectories.

The paper is structured as follows: Section 1 provides an introduction and describes the estimand for the paper. Section 2 describes the dataset, highlighting the variables used and their distributions. Section 3 provides an overview of the logistic regression model and its detailed explanations. Section 4 presents the visualized results, highlighting variability in vocabulary acquisition by age and word categories, as well as predicted probabilities. Finally, Section 5 delves into the findings, explores their implications, and outlines the limitations and future research directions. Appendix - A includes in-depth exploration and evaluation on data collection methods. Appendix - B contains the model diagnostics and accuracy evaluation. The data for this study is sourced from Braginsky (2024) and CDI (2024), with the entire analysis conducted in R Core Team (2023).

2 Data

2.1 Overview

The original dataset was obtained from CDI (2024) and downloaded using the R package Braginsky (2024). After a thorough cleaning process such as grouping related items and removing missing values, the analysis focuses on key variables, including category, age, comprehension, production and is_norming. These variables form the foundation of the analysis dataset, with an overview provided in Table 1.

Table 1: Cleaned Word Bank Dataset

Language	Age	Is_Norming	Broad_Category	Production	High_Vocabulary
English (American)	25	FALSE	Sensory Words	658	1
English (American)	26	FALSE	Sensory Words	552	1
English (American)	24	FALSE	Sensory Words	504	1
English (American)	26	FALSE	Sensory Words	272	0
English (American)	24	FALSE	Sensory Words	350	1
English (American)	25	FALSE	Sensory Words	580	1
English (American)	22	FALSE	Sensory Words	351	1
English (American)	24	FALSE	Sensory Words	310	1
English (American)	25	FALSE	Sensory Words	257	0

Table 1: Cleaned Word Bank Dataset

Language	Age	Is_Norming	Broad_Category	Production	High_Vocabulary
English (American)	26	FALSE	Sensory Words	188	0

2.2 Measurement

The objective of this study’s measurement approach is to transform raw parental reports into reliable indicators of children’s vocabulary acquisition patterns. The data is sourced from CDI (2024), a widely used tool that collects detailed information on children’s vocabulary comprehension and production through structured parental surveys. These surveys allow parents to report their child’s understanding and usage of specific words, which are categorized into lexical groups such as nouns, verbs, and adjectives. This structured framework provides the foundation for creating the dependent and independent variables used in the analysis.

- **Standardized Structure:** The CDI uses predefined response categories to minimize reporting ambiguity and ensure consistency across participants. While this format reduces variability in responses, it may not fully capture the nuances of vocabulary acquisition.
- **Words Categorization:** Vocabulary items are grouped into distinct lexical categories, such as Function Words, Sensory Words, and Adjectives, enabling a detailed examination of how different types of vocabulary are acquired.
- **Norming Sample Inclusion:** To improve validity, the dataset includes a subset of children from norming groups, serving as benchmarks for language development. While these groups provide a standard for comparison, they may not fully represent the diversity of the broader population.
- **Bias Mitigation:** The CDI’s structured responses and norming benchmarks help reduce bias but cannot fully eliminate inaccuracies from parental reporting. Parents may overestimate or underestimate their child’s abilities due to subjective perceptions or limited observations, which remains a limitation of the self-reported data.
- **Missing Data Handling:** Observations with incomplete or invalid responses were excluded from the analysis to maintain the integrity and reliability of the dataset.
- **Temporal Limitations:** Temporal Snapshot Limitation: The CDI captures vocabulary development at discrete age points, which may not fully reflect rapid, short-term changes in children’s language acquisition.

2.3 Outcome Variable

2.3.1 High Vocabulary Score

The outcome variable in this study, High Vocabulary Score, is a binary indicator designed to identify individuals with advanced vocabulary. This variable is derived from two key measures:

1. Comprehension: This variable represents the ability to understand words and phrases, reflecting the receptive language skills of individuals. Comprehension scores are numerical and vary across the dataset.
2. Production: This variable captures the ability to produce words, reflecting expressive language skills. Like comprehension, production scores are numerical and provide the standard into verbal articulation capabilities.
3. High Vocabulary Score: It is calculated using the average of comprehension and production scores for each individual. This average is represented as:
$$\text{prod_comp_mean} = \frac{\text{Comprehension} + \text{Production}}{2}$$

Referring to Taintor and LaMarr (2023) and exploratory data analysis, I investigated how productive vocabulary size evolves in children aged 16 to 30 months. Observing the 50th percentile, which approximates 300 words, I decided to create a binary variable using this threshold. This choice ensures the threshold reflects a typical vocabulary level for children aged 16 to 30 months, avoiding the influence of outliers. Using the 50th percentile also aligns with the goal of distinguishing between children with relatively advanced vocabulary (above the median) and those with less developed vocabulary (below the median), making it an effective and balanced benchmark for binary classification.

To classify individuals, a threshold value of 300 is applied to `prod_comp_mean`:

- Individuals with `prod_comp_mean > 300` are classified as having a high vocabulary score (outcome = 1).
- Those with `prod_comp_mean <= 300` are classified as not having a high vocabulary score (outcome = 0).

This method accounts for both receptive (comprehension) and expressive (production) skills in defining advanced vocabulary. A threshold of 300 provides a clear distinction between individuals with high and low vocabulary abilities. The High Vocabulary Score is used as the dependent variable in the subsequent model setup. Its binary nature makes it well-suited for analysis using a binomial family distribution and for estimating factors influencing advanced vocabulary acquisition.

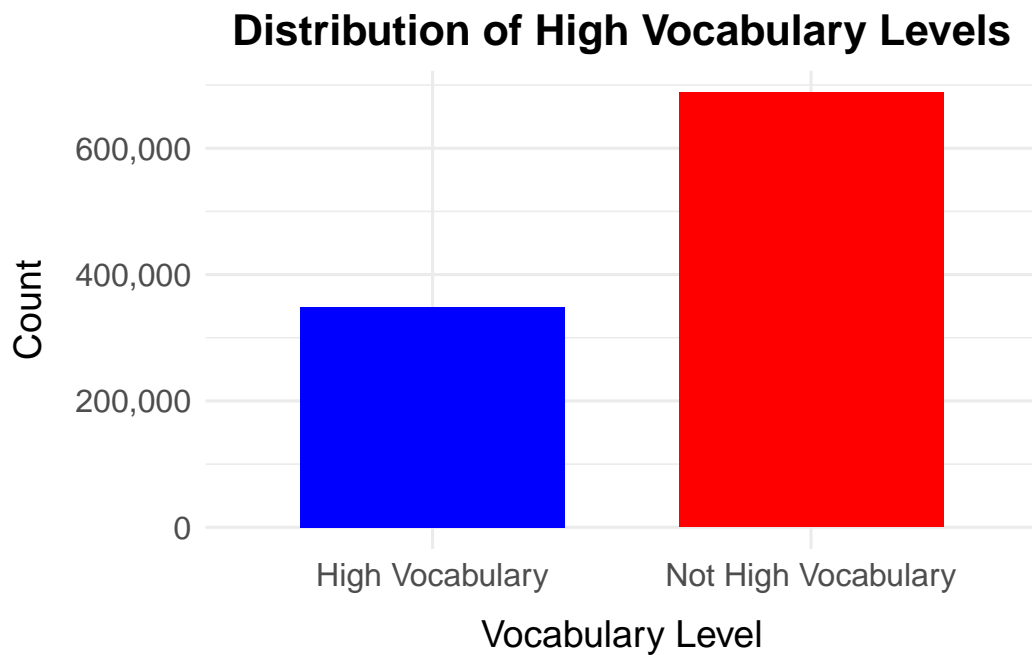


Figure 1: Distribution of high vocabulary levels among children aged 16–30 months. The chart shows a larger proportion of children classified as “Not High Vocabulary” compared to those with “High Vocabulary”.

2.4 Predictor Variables

2.4.1 Age

Age plays an important role in understanding vocabulary acquisition patterns in children, as it reflects developmental progress over time. As shown in Figure 2, the dataset reveals a concentration of participants between 24 and 30 months, a critical period during which children typically experience substantial gains in vocabulary comprehension and production. This higher representation is likely driven by a research emphasis on developmental milestones that characterize this stage, making the dataset particularly valuable for studying advanced vocabulary acquisition.

Distinct peaks in the dataset, such as at 25 and 30 months, may indicate intentional focus points for testing or align with developmental benchmarks tied to standardized assessments like the MacArthur-Bates Communicative Development Inventories (CDI). However, this uneven distribution also highlights the challenges of assessing vocabulary development in younger children, where verbal communication is less developed, and parental reporting is more variable. These disparities underscore the importance of standardizing age in statistical models to ensure accurate analysis and interpretation, while also exercising caution when generalizing findings to underrepresented age groups.

2.4.2 Word Category

The study categorizes words into broad lexical groups to analyze patterns of vocabulary acquisition, as depicted in Figure 3. These categories include Activities, Adjectives, Function Words, Living Things, Objects, Places, Sensory Words, and Verbs, with further subdivision of nouns into specific groups like Living Things, Objects, and Places to capture distinct trends. Function Words, encompassing pronouns and question words, reflect early grammatical development, while Verbs and Adjectives denote actions and descriptive expressions, both of which are critical for sentence construction and linguistic progression.

The distribution of word categories, shown in Figure 3, reveals notable differences in frequency. Objects form the largest category, reflecting the prominence of concrete and tangible items in children’s early vocabulary, as these are easier to recognize and recall. Verbs and Living Things follow as essential components of communication but appear less frequently. Conversely, Sensory Words and Activities are underrepresented, likely due to their context-specific and specialized nature. This variability emphasizes the role of concrete and functional words in early language development while also identifying potential gaps in underrepresented lexical groups.

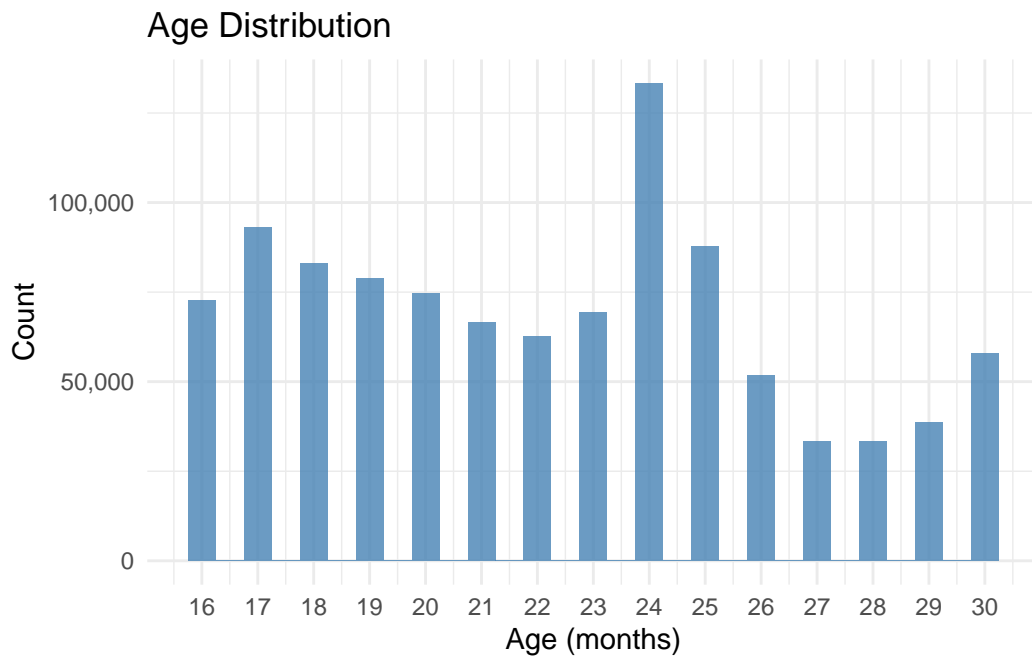


Figure 2: The graph visualizes the age distribution of children in the dataset, starting from 16 months. It shows a varying number of participants across age groups, with peaks around certain ages, such as 25 months, and a notable decline at the youngest and oldest ranges. This distribution highlights the dataset’s emphasis on certain developmental periods, which may reflect the focus of assessments or participant availability.

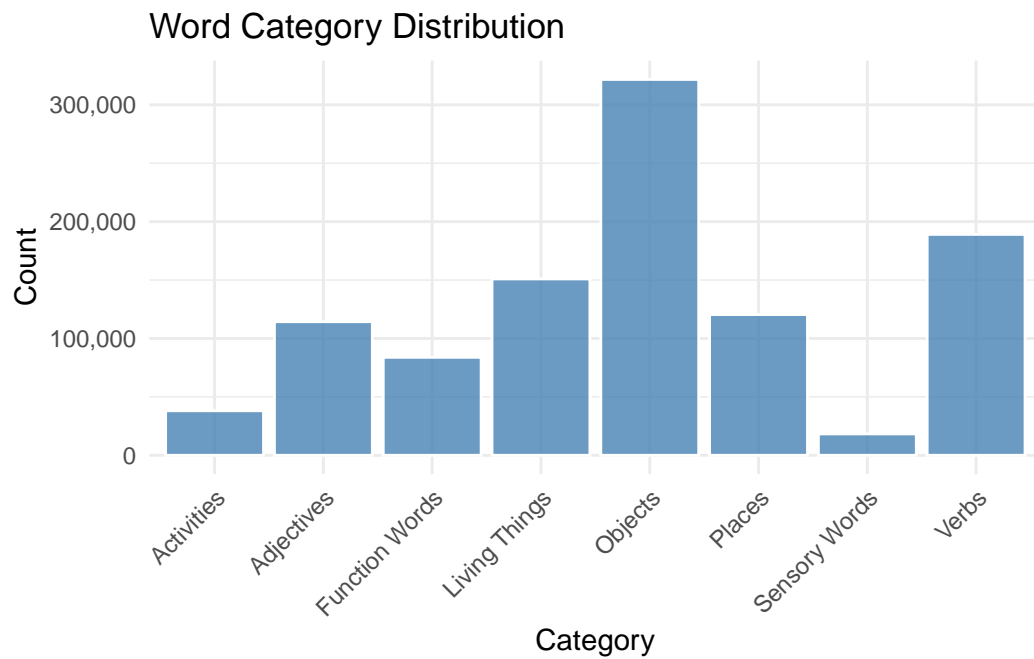


Figure 3: The figure shows objects dominate the vocabulary, reflecting an emphasis on concrete and tangible terms, while categories like Sensory Words and Activities are less frequently recorded.

2.4.3 Norming Status

The `is_norming` variable classifies children into two groups: norming and non-norming. The norming group serves as a standardized sample, providing benchmarks for assessing vocabulary development and enabling consistent comparisons across the dataset. Although the dataset is predominantly composed of non-norming children, norming data is critical for calibrating and interpreting learning patterns, particularly in early language acquisition studies or education.

This variable is included in the logistic regression model to account for systematic differences between the groups. By doing so, the model ensures that variations in the likelihood of achieving a high vocabulary score are not confounded by differences in group composition or assessment protocols. Despite the imbalance, this approach allows the model to differentiate patterns arising from population diversity versus those influenced by structured norming samples, as shown in Table 2.

Table 2: Summary of Norming Status in the Dataset

Norming Status	Count
FALSE	1031560
TRUE	5440

3 Model

To show the relationship between children’s vocabulary acquisition and their demographic and linguistic characteristics, a logistic regression model is constructed in this section. By examining key demographic and linguistic predictors, we aim to identify how characteristics like age, norming status, and word categories influence vocabulary development. The dependent variable, `high_vocabulary`, is a binary outcome indicating whether a child’s average production and comprehension score (denoted as `prod_comp_mean`) exceeds 300. This threshold was chosen to distinguish children with relatively advanced vocabulary levels. More background details and diagnostics are included in Appendix - B.

3.1 Model Overview

- **High Vocabulary:** A binary indicator where 1 represents a high vocabulary score (combined comprehension and production > 300), and 0 otherwise.
- **Scaled Age (`age_scaled`):** The child’s age, standardized to reflect changes per standard deviation. Standardization aids in interpretability and ensures numerical stability.

- Norming Status (is_norming): A binary indicator denoting whether a child is part of the norming dataset (TRUE) or not (FALSE). This variable accounts for potential differences in data collection or assessment protocols.
- Broad Category (broad_category): A categorical variable grouping words into lexical categories, such as adjectives, verbs, and nouns. The reference category for comparison is Function Words.

The model is specified as:

$$\log \left(\frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 \cdot \text{age_scaled}_i + \beta_2 \cdot \text{is_normingTRUE}_i \quad (1)$$

$$+ \beta_3 \cdot \text{broad_categoryAdjectives}_i \quad (2)$$

$$+ \beta_4 \cdot \text{broad_categoryFunction_Words}_i \quad (3)$$

$$+ \beta_5 \cdot \text{broad_categoryLiving_Things}_i \quad (4)$$

$$+ \beta_6 \cdot \text{broad_categoryObjects}_i \quad (5)$$

$$+ \beta_7 \cdot \text{broad_categoryPlaces}_i \quad (6)$$

$$+ \beta_8 \cdot \text{broad_categorySensory_Words}_i \quad (7)$$

$$+ \beta_9 \cdot \text{broad_categoryVerbs}_i \quad (8)$$

Where: - p_i represents the probability that child i has a high vocabulary score - β_0 is the intercept, capturing the baseline log-odds when all predictors are at their reference or mean levels - β_1 : Effect of age (standardized) - β_2 : The effect of whether the individual belongs to the norming group - $\beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9$: The effects of being in the respective broad word categories (nouns, function words, or verbs), compared to the reference category (likely “adjectives”).

3.2 Model Assumptions

- Linearity of the Logit: The model assumes a linear relationship between the log-odds of the outcome (high vocabulary proficiency) and the independent variables. For instance, the standardized age variable (age_scaled) ensures that for every one standard deviation increase in age, the log-odds of achieving high vocabulary change by a constant amount. Standardization centers the variable around zero and scales it to have a standard deviation of one, improving both interpretability and adherence to the linearity assumption.
- Independence of Observations: Each observation corresponds to a unique child, and the data assumes no repeated measures or nested structures (e.g., grouping by classrooms or schools). This independence ensures the validity of the logistic regression framework. If dependencies, such as repeated measures or clustering, were present, a mixed-effects logistic regression or other hierarchical modeling techniques would be required.

- **Categorical Variable Encoding:** The categorical variable `broad_category` (e.g., “Adjectives,” “Verbs,” “Living Things”) is encoded using sum contrasts. This ensures that each coefficient reflects the deviation of a given category from the overall mean effect across all categories. For example, the coefficient for “Verbs” represents the difference in log-odds for this category compared to the average log-odds across all categories. Sum contrasts further allow the model’s intercept to represent the overall mean effect when all predictors are at their reference or mean levels, facilitating clear interpretation.
- **Binary Nature of the Outcome:** The dependent variable, `high_vocabulary`, is binary (1 = high vocabulary, 0 = not high vocabulary). Logistic regression is appropriate for modeling binary outcomes, as it assumes the binomial distribution of the data, which aligns with the structure of the outcome variable.
- **No Perfect Multicollinearity:** The predictors are assumed to be non-perfectly correlated. High multicollinearity would lead to unreliable coefficient estimates and obscure the individual effects of predictors. Standardizing continuous predictors (e.g., age) and using appropriate encoding for categorical variables help minimize this risk and ensure stable model estimation.

3.3 Interpretation of Coefficients

The logistic regression coefficient (β) represent the change in the log-odds of achieving high vocabulary proficiency for a one-unit change in the respective predictor variable, holding all other variables constant.

- **Intercept (β_0):** Represents the log-odds of high vocabulary proficiency when all predictors are at their reference levels. If $\beta_0 > 0$, the baseline odds of high vocabulary are greater than 50%. Referring to the modelsummary in the Section B, we know that the baseline odds of high vocabulary proficiency are less than 50%.
- **Scaled Age (β_1):** For each one standard deviation increase in age, the log-odds of high vocabulary increase by β_1 . If $\beta_1 = 0.5$, then $\exp(0.5) \approx 1.65$, meaning the odds increase by 65% for every one standard deviation increase in age.
- **Norming Status (β_2):** Indicates the effect of belonging to the norming group, a positive β_2 suggests higher odds of high vocabulary compared to non-norming children. If $\beta_2 = 0.1$, then $\exp(0.1) \approx 1.11$, meaning norming group children have 22% higher odds of high vocabulary.
- **Broad Category ($\beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9$):** The coefficients for the broad word categories (e.g., Function Words, Living Things) represent their effect on the log-odds of high vocabulary compared to the reference category (Adjectives). A positive β_k indicates higher odds compared to Adjectives. For example, if $\beta_4 = 0.3$ (Function Words), then $\exp(0.3) \approx 1.35$, meaning Function Words increase the odds by 35% compared to Adjectives. A negative β_k implies lower odds. For instance, if $\beta_7 = -0.008$ (Places), then $\exp(-0.008) \approx 0.992$ shows a minor decrease in odds for words in the Places category.

3.4 Model Justification

Logistic regression is well-suited for this study, given its ability to model binary outcomes like high vocabulary proficiency (1 = high vocabulary, 0 = not high vocabulary) while maintaining interpretability. By constraining predicted probabilities between 0 and 1, logistic regression ensures meaningful predictions. Its coefficients offer clear investigation into the magnitude and direction of effects, enabling an understanding of how predictors like age and norming status influence the likelihood of high vocabulary acquisition. For instance, odds ratios derived from the model allow straightforward interpretation of the impact of each variable, such as the increased likelihood of high vocabulary with a one-standard-deviation increase in age.

Although more advanced machine learning models like random forests or neural networks could provide slight improvements in predictive accuracy, these approaches lack the transparency needed to understand the underlying relationships between predictors and outcomes. Given this study’s focus on uncovering developmental patterns rather than solely optimizing prediction accuracy, logistic regression offers the necessary balance between interpretability and performance. Moreover, the dataset size and structure favor logistic regression, which is less prone to overfitting than more complex models that often require larger datasets to generalize effectively.

To ensure model robustness, the dataset was split into training and testing subsets to validate the model and minimize overfitting. Additionally, predictors like age were standardized to ensure comparability and prevent dominance by variables with larger numerical ranges. Categorical variables, such as word categories, were encoded using sum contrasts, allowing meaningful comparisons and ensuring that coefficients reflect deviations from the overall mean effect. These steps, combined with logistic regression’s simplicity and explanatory power, make it an optimal choice for investigating vocabulary acquisition patterns.

4 Results

4.1 Variability in Production Vocabulary

Figure 4 visualizes the relationship between age and production vocabulary scores, focusing on different percentiles of the distribution. The scatterplot shows individual production scores as gray dots, while overlaid lines represent percentiles (10th, 25th, 50th, 75th, and 90th), capturing central tendencies and variability across ages. The 50th percentile (median) line provides a benchmark for the typical vocabulary production score at each age, whereas the 10th and 90th percentiles outline the lower and upper ranges of vocabulary development. The gradual upward trend of the median line reflects consistent growth in production vocabulary as children age, with a widening gap between the percentiles at later ages. This widening suggests increasing variability in vocabulary acquisition, with some children advancing much faster than others in production abilities.

The data indicates that children in the 90th percentile acquire vocabulary at a significantly faster rate than their peers, as evidenced by the steeper slope of the topmost line. Conversely, the 10th and 25th percentiles show more gradual, stable growth, suggesting slower development for children in these groups. The broader range of scores at older ages emphasizes the heterogeneity of developmental trajectories, with some children reaching vocabulary sizes substantially larger than the median while others remain below average. These findings underscore the diversity in early language acquisition and highlight the importance of considering individual differences when evaluating children’s vocabulary development.

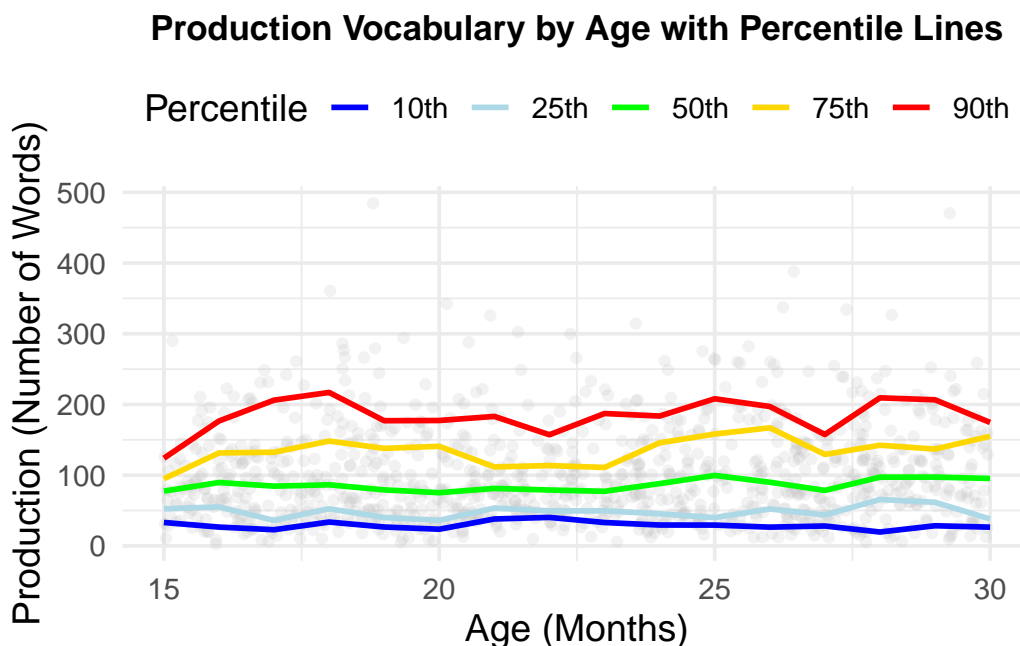


Figure 4: Production Vocabulary by Age with Percentile Lines. The graph illustrates the production scores of children across different ages (in months). Individual data points (gray dots) represent raw production scores for each participant. Colored lines correspond to standardized percentiles—10th (blue), 25th (light blue), 50th (green), 75th (yellow), and 90th (red)—showing trends in vocabulary production distribution over time.

4.2 Median Vocabulary Size Change by Age

Figure 5 illustrates the median comprehension vocabulary scores across different ages, focusing on the central tendency of children’s comprehension development between 15 and 30 months. The gray dots represent individual data points, capturing the variability in comprehension scores, while the blue line highlights the median score for each age group. The graph shows a

clear upward trajectory, with median comprehension steadily increasing with age, particularly after 18 months. This suggests a critical developmental period between 18 and 30 months during which children experience significant growth in comprehension vocabulary. The density and spread of gray points around the median line indicate individual variability, emphasizing that while the general trend is one of growth, some children exhibit slower or faster development compared to their peers. The visualization underscores the importance of age as a determinant of vocabulary comprehension while highlighting the diverse range of learning patterns among children.

Between 15 and 18 months, the median comprehension score remains relatively stable, indicating slower growth in vocabulary during early stages of language acquisition. A noticeable increase in vocabulary size is observed after 18 months, suggesting that children begin to acquire words more rapidly as their cognitive and linguistic abilities develop. The most significant growth occurs between 24 and 30 months, where the median comprehension score consistently rises. This period aligns with critical developmental milestones, such as the expansion of receptive language and comprehension skills. Around 30 months, the upward slope of the median line begins to level off slightly, suggesting that comprehension growth may slow down or stabilize as children approach the end of the observed range.

The use of the median instead of the mean ensures that the central trend is not skewed by outliers (e.g., extremely high or low comprehension scores). This choice provides a robust summary of comprehension at each age, especially in datasets with large variability or non-normal distributions. These findings highlight the critical window between 21 and 26 months for comprehension vocabulary growth. Interventions or language exposure strategies during this period may be particularly effective in enhancing language development. The observed variability suggests that individual-level factors (e.g., family environment, exposure to language) play a significant role in shaping comprehension scores, warranting further investigation into these influences.

4.3 Prediction for the Probability of High Vocabulary Level

4.4 Predicted Probabilities by Age

Figure 7 illustrates the relationship between predicted probabilities of achieving high vocabulary and age, with percentile trends (10th, 25th, 50th, 75th, and 90th percentiles) overlaid to highlight the variability in predictions. Based on the resample of the test data, the scatter-plot points represent individual predicted probabilities, while the percentile lines depict the progression of predictions across age groups.

The median predicted probability (50th percentile, solid blue line) steadily increases with age, reflecting the model's growing confidence in high vocabulary acquisition as children get older. The 10th and 25th percentiles (dashed and dotted red/purple lines) remain relatively

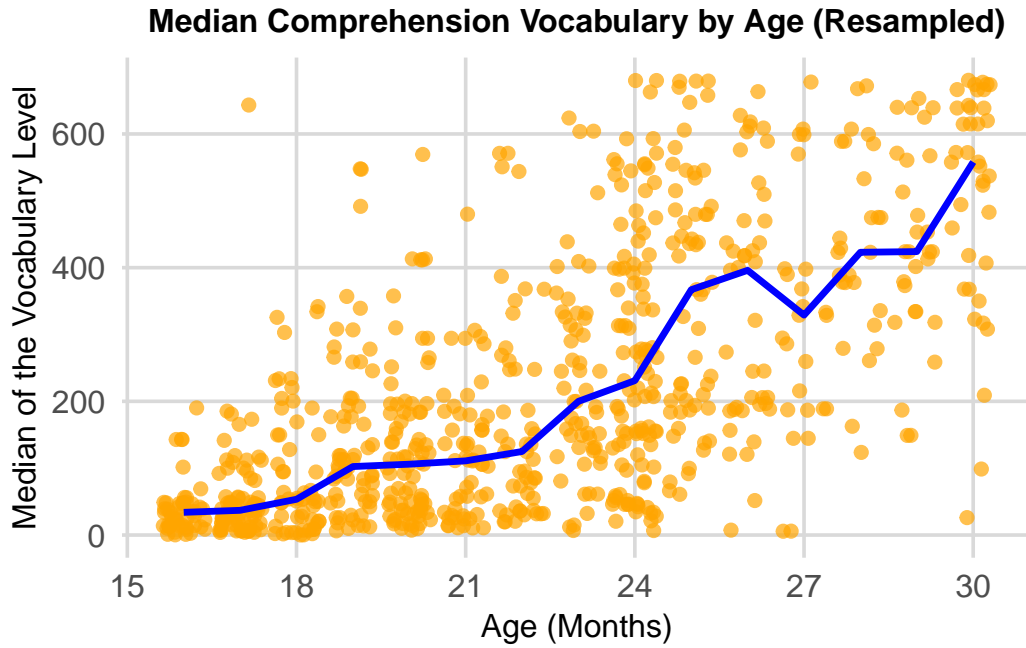


Figure 5: This figure illustrates the relationship between age (in months) and the median comprehension vocabulary level. Each orange point represents a resampled data point, while the blue line depicts the median vocabulary level at each age. The trend highlights a steady increase in vocabulary comprehension as children age, with noticeable variability across individual data points.

age	broad_category	is_norming	mean_predicted_prob
Min. :16.00	Activities :22	Mode :logical	Min. :0.02989
1st Qu.:20.00	Adjectives :22	FALSE:120	1st Qu.:0.18404
Median :23.00	Function Words:22	TRUE :56	Median :0.54167
Mean :23.27	Living Things :22	NA	Mean :0.49883
3rd Qu.:27.00	Objects :22	NA	3rd Qu.:0.78088
Max. :30.00	Places :22	NA	Max. :0.97843
NA	(Other) :44	NA	NA

Figure 6: How the probability of high vocabulary varies with age by aggregating the predictions for each age group

low at younger ages but show a notable rise after 20 months, suggesting greater variability in predictions among younger children. In contrast, the 75th and 90th percentiles (dashed and dotted purple/red lines) start higher and climb more sharply, indicating that some children exhibit advanced vocabulary skills even at younger ages.

Overall, the increasing spread between the percentile lines with age highlights a broader range of vocabulary acquisition patterns as children develop. This visualization underscores the model's ability to account for individual differences in learning trajectories while confirming the strong association between age and predicted probability of high vocabulary.

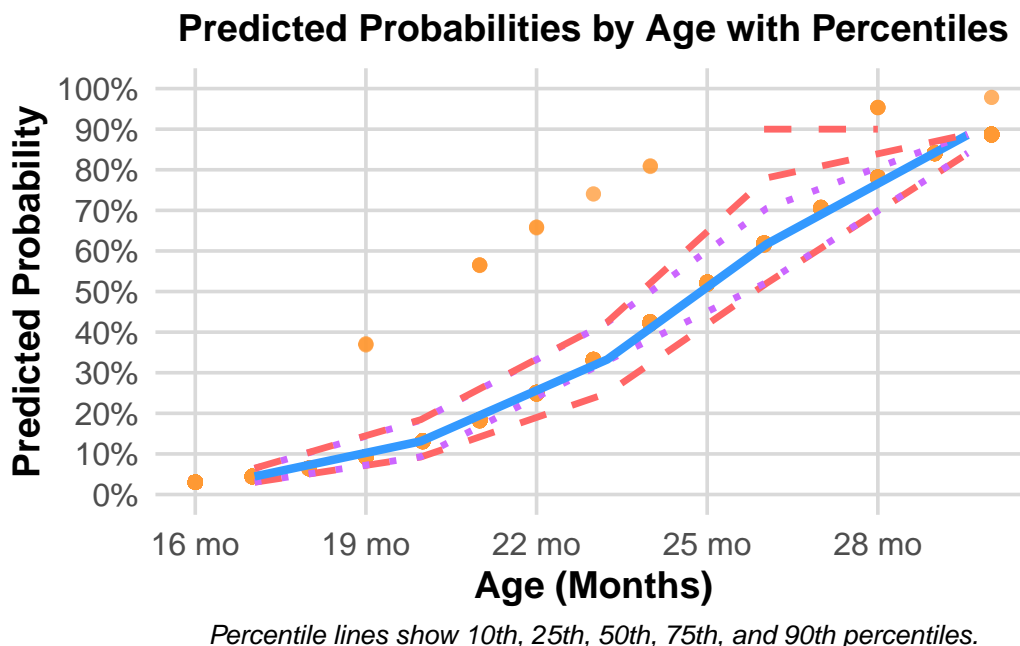


Figure 7

4.5 Distribution of Predicted Probabilities by Word Category

The distribution of predicted probabilities for achieving high vocabulary varies across word categories, as shown in Figure 8. Categories such as Function Words, Living Things, and Objects display a concentration of higher predicted probabilities, indicating consistent acquisition patterns and the foundational role of these words in early communication. In contrast, categories like Sensory Words and Verbs show a higher density of lower probabilities, reflecting their later emergence in language development and the contextual complexity they require.

Categories such as Adjectives and Places exhibit broader distributions, spanning a wider range of predicted probabilities. This variability highlights differences in children's exposure to and

use of these words, suggesting that external factors like linguistic environments may play a significant role. Overall, these results emphasize how word categories shape the model’s predictions and highlight the importance of addressing variability in complex word types for targeted interventions.

5 Discussion

5.1 Age and Developmental Dynamics

The role of age in vocabulary acquisition extends beyond a simple progression of linguistic capabilities—it reflects the interplay of rapid cognitive, social, and linguistic development that unfolds during early childhood. As shown in the results, the period between 18 and 30 months is characterized by an accelerated growth in vocabulary size, with the steepest increases observed in comprehension and production scores. This critical window not only underscores the biological readiness of children to absorb language but also highlights how external factors, such as parental interaction and linguistic exposure, amplify the trajectory of language growth.

Interestingly, the variability in acquisition patterns suggests that age alone does not act uniformly across all children. For instance, the widening gap in vocabulary sizes at older ages indicates that while some children reach advanced levels of vocabulary acquisition, others lag behind. This divergence raises questions about the influence of environmental and contextual factors—such as the quality and quantity of language input, socio-economic conditions, or bilingualism—that shape individual trajectories. Moreover, the plateau observed in vocabulary comprehension beyond 30 months suggests a potential shift in focus from foundational vocabulary to more nuanced, context-dependent language learning. These findings highlight the need for age-specific interventions that cater to both foundational vocabulary during early months and more complex linguistic skills as children grow older.

5.2 Word Categories: Patterns and Variability

The analysis of word categories reveals distinct patterns in vocabulary acquisition, highlighting differences in cognitive accessibility and linguistic exposure. Foundational categories such as Function Words, Living Things, and Objects show consistently higher predicted probabilities, emphasizing their central role in early communication. These categories, linked to tangible items and routine interactions, form the basis for early linguistic development and are acquired more universally across children.

On the other hand, categories like Sensory Words and Adjectives display greater variability in predicted probabilities, reflecting their reliance on contextual understanding and experiential learning. Sensory words often require abstract reasoning or sensory experiences that develop

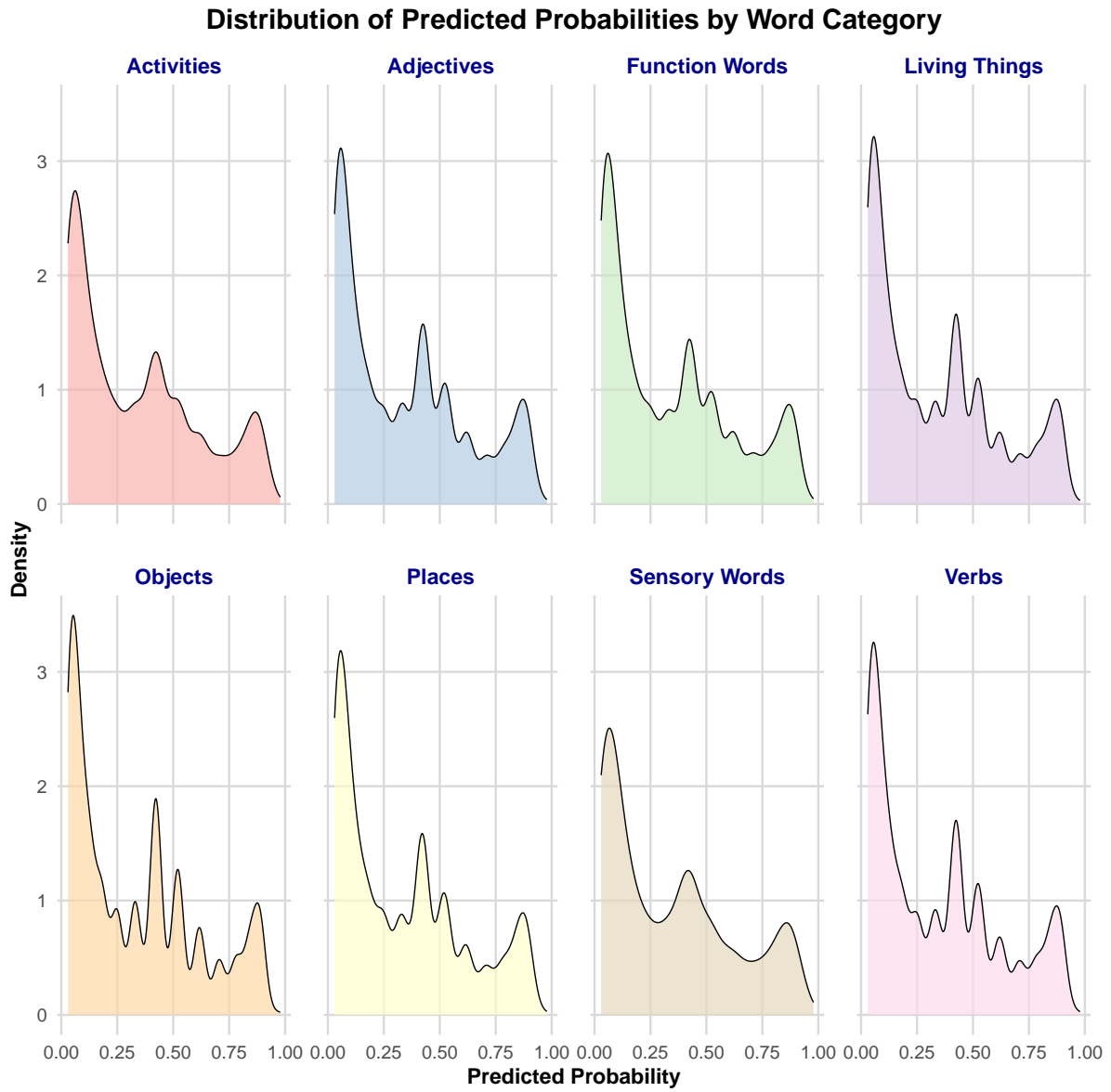


Figure 8: This figure displays the density distributions of predicted probabilities for achieving high vocabulary across different word categories. Categories such as Function Words and Living Things show concentrated higher probabilities, reflecting consistent acquisition patterns, while categories like Verbs and Sensory Words exhibit broader distributions and lower probabilities.

later, while adjectives demand an ability to describe and compare attributes, which may not emerge until later stages. These patterns suggest opportunities to support the acquisition of less frequently encountered categories by incorporating them into everyday activities, such as interactive storytelling or guided play.

Verbs, occupying a middle ground, show moderate variability due to their dependence on both cognitive development and frequent reinforcement through daily interactions. These words often appear in action-oriented communication, such as instructions or play, making them integral but slightly more challenging to master. Addressing these category-specific challenges can help promote more balanced vocabulary development and support children in acquiring both foundational and complex word types.

5.3 Implications for Early Childhood Education

The findings of this study offer valuable contributions to early childhood education by emphasizing the importance of targeted strategies in fostering balanced vocabulary development. Early vocabulary acquisition, as demonstrated, is not uniform across lexical categories. Foundational categories such as Function Words and Objects are consistently acquired and should be reinforced through structured linguistic activities, while categories like Sensory Words and Adjectives require more deliberate and context-specific teaching strategies due to their variability and reliance on experiential learning.

Educators and caregivers can use these insights to design age-appropriate learning environments that bridge gaps in underdeveloped word categories. For instance, integrating sensory-rich experiences, such as interactive play or descriptive storytelling, can provide children with exposure to Sensory Words and Adjectives, enhancing their cognitive and linguistic skills. Similarly, action-based activities that encourage the use of Verbs can be incorporated into daily routines, allowing children to connect language with physical movement and actions.

Additionally, the study highlights the critical role of age-specific interventions. For children aged 16–24 months, the focus should be on expanding foundational vocabulary to ensure a strong linguistic base. For those aged 25–30 months, introducing more complex and abstract words can help develop higher-order language skills. Structured programs, such as those integrating books, games, and parent-child interaction, can be tailored to these developmental needs.

Lastly, the disparities observed between norming and non-norming groups underscore the importance of equitable access to high-quality linguistic environments. Early childhood education programs should aim to provide diverse linguistic exposure across socio-economic and cultural contexts, ensuring all children have the opportunity to reach their vocabulary potential. By applying these findings, educators and policymakers can better support children’s language development, fostering their cognitive and academic success.

5.4 Limitations and Future Directions

Song et al. (2015) also point out in their study on 264 typically developing Chinese children (145 boys and 119 girls) were included from a longitudinal study of language. While the findings offer valuable directions, several limitations should be acknowledged.

This study sheds light on key aspects of vocabulary acquisition, but certain limitations must be acknowledged. First, the reliance on parental reports, as collected through the MacArthur-Bates Communicative Development Inventories (CDI), introduces potential biases. Parents may overestimate or underestimate their child’s abilities due to memory limitations, subjective interpretation, or social desirability. Such biases could affect the accuracy of the data, particularly for categories requiring nuanced understanding, like Sensory Words. To mitigate these limitations, future research should incorporate complementary methods, such as direct observational studies or experimental assessments, to provide more reliable and objective measurements of children’s vocabulary development.

The cross-sectional nature of the dataset also limits the ability to capture individual developmental trajectories over time. While cross-sectional data offers a snapshot of vocabulary acquisition, it cannot account for within-child variability or the dynamic progression of language learning. Longitudinal studies are essential for understanding how vocabulary grows and evolves across different ages and contexts. By tracking the same children over extended periods, researchers could uncover mechanisms driving individual differences in learning rates, identify critical intervention windows, and observe the long-term effects of early linguistic exposure.

Another limitation lies in the dataset’s word category representation. While foundational categories like Function Words and Objects are well-represented, others, such as Sensory Words and Adjectives, are underrepresented, limiting the analysis of these more complex and context-dependent types. Future datasets should strive for a more balanced representation of word categories to explore their unique developmental trajectories. Additionally, integrating contextual factors such as socio-economic status, bilingualism, and parental language practices into future studies could provide a deeper understanding of how these variables influence vocabulary growth. These enhancements would allow for more comprehensive models that account for environmental diversity and linguistic richness.

Addressing these limitations through longitudinal studies, enriched datasets, and mixed-methods approaches can deepen our understanding of early vocabulary development. By incorporating diverse populations, balanced lexical categories, and multi-dimensional data, future research can offer actionable insights to support tailored educational interventions and equitable language development strategies for children from all backgrounds.

(Firstly, the reliance on parental reports introduces potential biases, including over- or underestimation of children’s abilities. Future studies could complement CDI data with observational or experimental measures to enhance reliability. Second, the cross-sectional nature of the data limits the ability to track individual developmental trajectories. Longitudinal

studies are needed to capture within-child variability and the dynamics of vocabulary growth over time. Familial factors and reading or language related cognitive skills were found to be associated with these developmental subgroups.

Additionally, the dataset we have now is limited on the word categories. Future data collection could focus more on the word types and provides more feasible analysis on the lexical category learning pattern. Besides, Future research should also explore the influence of environmental and contextual factors, such as language exposure, educational interventions, and socio-economic status, on vocabulary acquisition. These factors could provide a more comprehensive understanding of the mechanisms underlying linguistic development.

The strongest developmental inferences can be made by the examination of longitudinal data, in which children’s individual development is measured multiple times using the same instrument. Unfortunately, relatively little of our CDI data comes from this type of repeated administration. There is a substantial amount of two-administration longitudinal data for several languages, but only a few have more than two observations for individual children. In general, this aspect of our data is a consequence of the fact that, for normative datasets, pure cross-sectional data collection is used to ensure statistical independence between datapoints. Thus, we must typically settle for using the large amount of available cross-sectional data to average out individual variability.)

A Appendix

A.1 Survey Design: The CDI Framework

The MacArthur-Bates Communicative Development Inventories (CDI) form the foundation of the Wordbank dataset used in this study. The CDI is a parent-report instrument that collects data on children’s vocabulary comprehension and production through structured surveys. This methodology allows for large-scale data collection across diverse populations, balancing cost-efficiency and practicality. However, as a parent-reported tool, the CDI is subject to several biases. For instance, parental interpretations of word comprehension or usage can vary, and factors like educational background, socio-economic status, and linguistic environment may influence the accuracy of responses.

The design of the CDI ensures comprehensive coverage of early vocabulary acquisition by categorizing words into broad lexical groups (e.g., Function Words, Verbs, Sensory Words). This categorization supports nuanced analysis but introduces challenges. For instance, frequently observed word types like Function Words may be more accurately reported than abstract categories like Sensory Words, leading to potential underestimation in less salient categories. These characteristics highlight the trade-off between the CDI’s extensive reach and the subjective variability in parental reporting.

A.2 Sampling Framework

The Braginsky (2024) dataset includes norming and non-norming samples, which together capture a broad spectrum of linguistic development. Norming samples are designed to represent a balanced, standardized population, serving as benchmarks for vocabulary acquisition. These data allow for cross-study comparisons and robust generalizations but may not fully capture real-world variability. In contrast, non-norming samples encompass a wider range of linguistic environments, including underrepresented groups. This dual approach helps investigate effects of contextual diversity on language acquisition but also introduces challenges in integrating findings from norming and non-norming subsets.

One critical consideration in observational datasets like Wordbank is the potential for sampling biases. For example, children from bilingual households or lower socio-economic backgrounds may be underrepresented in the norming sample, leading to an incomplete picture of vocabulary acquisition patterns. Future surveys could benefit from targeted recruitment strategies to ensure more comprehensive sampling of diverse populations.

A.3 Observational Nature of the Data

The Wordbank dataset is observational, capturing vocabulary acquisition as it naturally occurs in diverse contexts. While this design enhances ecological validity, it limits causal inference. Factors such as parental education, linguistic exposure, and cultural norms may confound the observed relationships between predictors (e.g., age, norming status) and vocabulary outcomes. For instance, children in norming samples might have greater exposure to structured learning environments, inflating their vocabulary scores compared to non-norming counterparts.

To address these limitations, simulation methods were applied to assess how sampling variability influences model estimates. Random subsamples were drawn from the data, and the logistic regression model was refitted to evaluate the consistency of predictor effects. Key predictors like age and norming status remained stable, confirming the robustness of the model. However, coefficients for categories like Sensory Words and Places exhibited higher variability, reflecting their smaller sample representation and the challenges of capturing abstract vocabulary through parent-reported data.

A.4 Connecting to the Literature

The CDI framework and its use in Wordbank align with best practices in developmental research, as outlined in studies from Mayor and Mani (2018). The test-retest reliability of the CDI supports its validity, although the variability in parental reporting—especially for comprehension items—has been noted. Advanced statistical techniques like Item Response Theory (IRT) have been employed to evaluate the discriminative power of individual CDI items, revealing strengths in measuring commonly acquired words and limitations for abstract

concepts. These findings inform the design of future surveys, emphasizing the need for complementary methods, such as direct assessments or longitudinal designs, to address the limitations of observational data.

A.5 Simulation and Recommendations

To further validate the findings, simulations could be expanded to incorporate potential sampling biases. For example, generating synthetic datasets that account for underrepresented groups (e.g., bilingual or low socio-economic households) could show how these populations might influence observed patterns. Additionally, calibration exercises—where parents undergo brief training on how to interpret survey items—could enhance the consistency of responses and reduce reporting variability.

This appendix underscores the importance of survey design, sampling methodology, and observational data analysis in understanding early vocabulary acquisition. By addressing inherent limitations and leveraging advanced methodologies, future research can refine the robustness and applicability of findings, ensuring a more inclusive representation of children’s linguistic development.

B Model details

B.1 Model Summary

B.2 Diagnostics

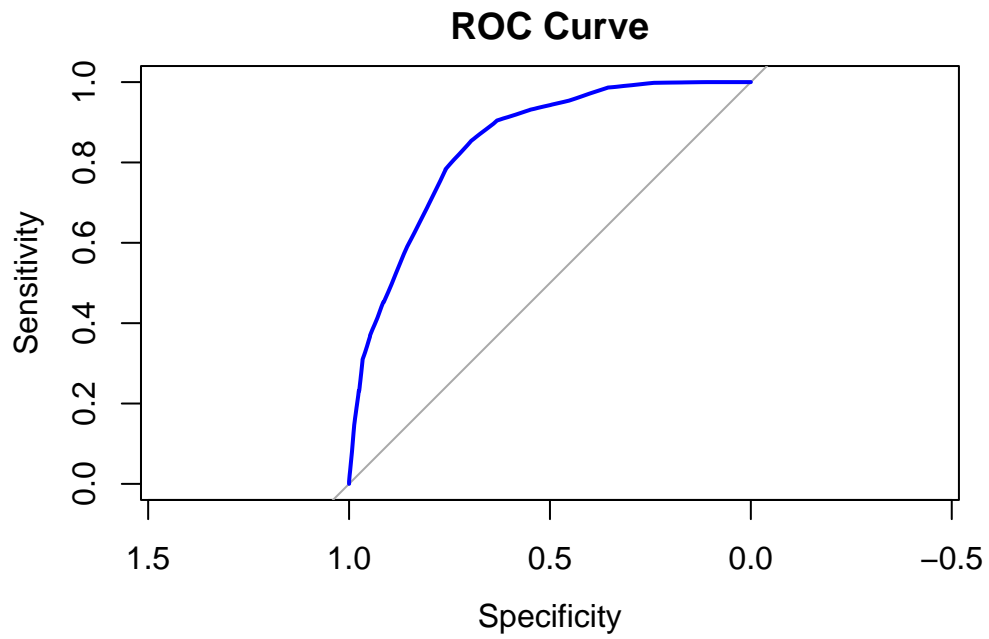
B.2.1 Confusion Matrix

Metric	Value
Accuracy	0.77
Sensitivity (Recall)	0.59
Specificity	0.86
Precision	0.67

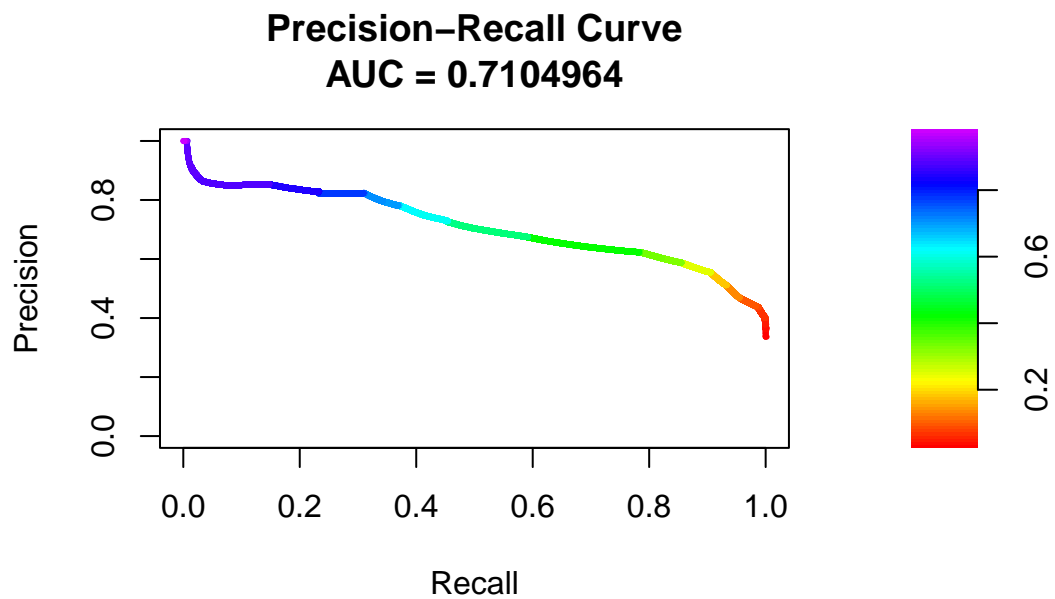
Table 4: Model Summary

	(1)
(Intercept)	−1.014 (0.015)
age_scaled	1.605 (0.004)
is_normingTRUE	1.749 (0.040)
broad_categoryAdjectives	0.006 (0.017)
broad_categoryFunction Words	0.005 (0.018)
broad_categoryLiving Things	0.005 (0.017)
broad_categoryObjects	−0.005 (0.016)
broad_categoryPlaces	−0.008 (0.017)
broad_categorySensory Words	−0.015 (0.026)
broad_categoryVerbs	0.004 (0.016)
Num.Obs.	829 600
AIC	754 890.6
BIC	755 006.9
Log.Lik.	−377 435.316
RMSE	0.39

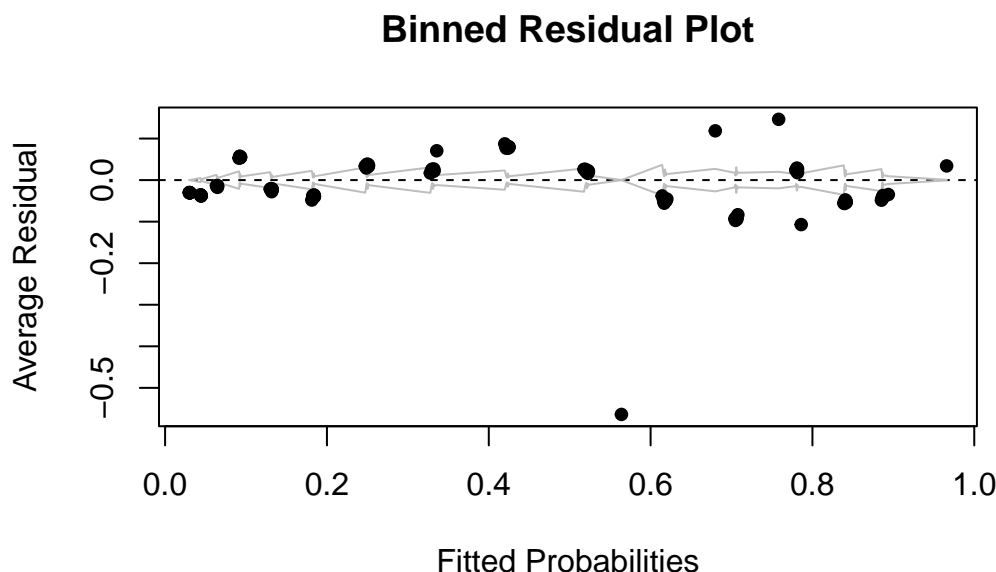
B.2.2 ROC Curve and AUC



B.2.3 Precision-Recall Curve



B.2.4 Binned Residual Plot



C Acknowledgements

This project was conducted under the help of OpenAI’s ChatGPT 4.0, which provided invaluable assistance in drafting and refining the paper. The analysis was conducted using a suite of packages from R Core Team (2023), which offered robust functionality for data manipulation, visualization, and storage. We extend our gratitude to the teams behind the Wickham et al. (2019), Wickham (2016), Wickham, Pedersen, and Seidel (2023), Wickham et al. (2023), Wickham, Vaughan, and Girlich (2024), Gelman and Su (2024), Arel-Bundock (2022) and Xie (2024) packages, whose tools were instrumental in streamlining the data cleaning, analysis, and graphing processes. Additionally, Richardson et al. (2024) played a critical role in efficient data handling and storage through Parquet files. Thanks to Robin et al. (2011) and Keilwagen, Grosse, and Grau (2014) for supporting model diagnostics in this study.

A special acknowledgment goes to the Braginsky (2024) and CDI (2024) team for providing the extensive dataset that forms the foundation of this research. Their contribution enabled a comprehensive exploration of vocabulary learning patterns in children. We are deeply grateful to the developers and maintainers of these open-source tools and datasets for their efforts in advancing research and accessibility in the data science community.

References

- Arel-Bundock, Vincent. 2022. “modelssummary: Data and Model Summaries in R.” *Journal of Statistical Software* 103 (1): 1–23. <https://doi.org/10.18637/jss.v103.i01>.
- Braginsky, Mika. 2024. *wordbankr: Accessing the Wordbank Database*. <https://github.com/langcog/wordbankr>.
- CDI, MacArthur-Bates Communicative Development Inventories. 2024. “MacArthur-Bates Communicative Development Inventories.” <https://mb-cdi.stanford.edu/>.
- Gelman, Andrew, and Yu-Sung Su. 2024. *arm: Data Analysis Using Regression and Multi-level/Hierarchical Models*. <https://CRAN.R-project.org/package=arm>.
- Keilwagen, Jens, Ivo Grosse, and Jan Grau. 2014. “Area Under Precision-Recall Curves for Weighted and Unweighted Data.” *PLOS ONE* 9 (3). <https://doi.org/10.1371/journal.pone.0092209>.
- Mayor, Julien, and Nivedita Mani. 2018. “A Short Version of the MacArthur–Bates Communicative Development Inventories with High Validity.” *Behavior Research Methods* 51 (October). <https://doi.org/10.3758/s13428-018-1146-0>.
- R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoş Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *arrow: Integration to ‘Apache’ ‘Arrow’*. <https://CRAN.R-project.org/package=arrow>.
- Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. “pROC: An Open-Source Package for r and s+ to Analyze and Compare ROC Curves.” *BMC Bioinformatics* 12: 77.
- Song, Suiping, Mengmeng Su, Chunyan Kang, Hongyun Liu, Yan Zhang, Catherine McBride-Chang, Twila Tardif, et al. 2015. “Tracing Children’s Vocabulary Development from Preschool Through the School-Age Years: An 8-Year Longitudinal Study.” *Developmental Science* 18 (1): 119–31. <https://doi.org/10.1111/desc.12190>.
- Taintor, Sue, and Wendy LaMarr. 2023. “Charting Language Growth in Infants and Toddlers.” LibreTexts. https://socialsci.libretexts.org/Bookshelves/Early_Childhood_Education/Infant_and_Toddler_Care_and_Development_%28Taintor_and_LaMarr%29/11%3A_Overview_of_Language_Development/11.09%3A_Charting_language_growth_in_infants_and_toddlers.
- Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *dplyr: A Grammar of Data Manipulation*. <https://CRAN.R-project.org/package=dplyr>.
- Wickham, Hadley, Thomas Lin Pedersen, and Dana Seidel. 2023. *scales: Scale Functions for*

Visualization. <https://CRAN.R-project.org/package=scales>.
Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2024. *tidyr: Tidy Messy Data*.
<https://CRAN.R-project.org/package=tidyr>.
Xie, Yihui. 2024. *knitr: A General-Purpose Package for Dynamic Report Generation in R*.
<https://yihui.org/knitr/>.