English Vocabulary Learning Pattern*

My subtitle if needed

Yongqi Liu

November 21, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

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

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2 descirbe the dataset in detail, and shows the distribution of variables.

The data used in this study comes from Braginsky (2024), and the whole paper is conducted and analyzed in R Core Team (2023)

2 Data

2.1 Overview

The original dataset was obtained from Braginsky (2024). After undergoing a thorough cleaning process—including grouping related items and removing missing values—the anal-

ysis focuses on the key variables: category, age, comprehension, production, is_norming, and broad_category. These variables form the foundation of the analysis dataset. An overview of the cleaned dataset is presented 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	0
English (American)	25	FALSE	Sensory Words	580	1
English (American)	22	FALSE	Sensory Words	351	1
English (American)	24	FALSE	Sensory Words	310	0
English (American)	25	FALSE	Sensory Words	257	0
English (American)	26	FALSE	Sensory Words	188	0

2.2 Measurement

2.2.1 Data Collection and Quality Control

The objective of measurement in this study is to translate raw parental reports into reliable indicators of vocabulary acquisition patterns in children. The data is derived from the MacArthur-Bates Communicative Development Inventories (CDI), a widely used tool that collects information on children's vocabulary comprehension and production through structured parental surveys. These surveys allow parents to report on their child's understanding and use of specific words, grouped into lexical categories such as nouns, verbs, and adjectives. The raw data collected through the CDI forms the basis for creating the study's dependent and independent variables.

```
-   Structured Response Formats: The CDI employs predefined response options for comprehensi
-   Lexical Categorization: Vocabulary items are grouped into meaningful lexical categories,
-   Norming Group Comparison: The inclusion of norming groups as benchmarks helps to ensure t
- Variable Standardization: Continuous variables, such as age, are standardized (e.g., scale
-   Bias Mitigation: By structuring responses and including norming benchmarks, the CDI minim
-   Missing Data Handling: Observations with incomplete or invalid responses were excluded fr
```

2.2.2 Reporting Bias

However, there are several considerations regarding the data collection process:

- Parental Reporting Bias: The reliance on parental reports introduces the potential for bias, including overestimation or underestimation of a child's abilities. This is inherent to self-reported data and can affect the accuracy of the results.
- Standardized Format and Structure: The CDI employs predefined response categories, which help to minimize ambiguity in reporting and ensure consistency across respondents. This structured approach mitigates some reporting variability but may not fully capture nuances in vocabulary acquisition.
- Norming Group Representation: To improve validity, a subset of children from norming groups is included as a benchmark for comparison. While useful, this raises concerns about whether the norming group adequately represents the population's diversity in language development.
- Temporal Limitations: The CDI data represents snapshots of vocabulary development at specific ages, which may not account for rapid changes or variations over time in a child's language acquisition process.

Despite its standardized structure, the CDI is subject to biases inherent in parental reporting, including Over or Underestimation Bias. Parents may unintentionally over- or underestimate their child's skills due to subjective perceptions or limited observations. Social Desirability Bias: Responses may be influenced by parents' desire to portray their child's language development favorably.

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 proficiency. 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. The High Vocabulary Score is calculated using the average of comprehension and production scores for each individual. This average is represented as: prod_comp_mean = $\frac{\text{Comprehension}+\text{Production}}{2}$

To classify individuals, a threshold value of 350 is applied to `prod_comp_mean`: - Individuals with `prod_comp_mean > 350` are classified as having a high vocabulary score (outcome = 1).

- Those with `prod_comp_mean <= 350` are classified as not having a high vocabulary score (outcome = 0).

This approach ensures that both receptive (comprehension) and expressive (production) skills are considered in defining advanced vocabulary. The threshold of 350 was chosen based on exploratory analysis of the dataset, reflecting a meaningful distinction between individuals with high and low vocabulary abilities. The High Vocabulary Score serves as the dependent variable in the following data analysis part. Its binary nature makes it suitable for modeling with a binomial family distribution, allowing for the estimation of factors that influence advanced vocabulary acquisition.

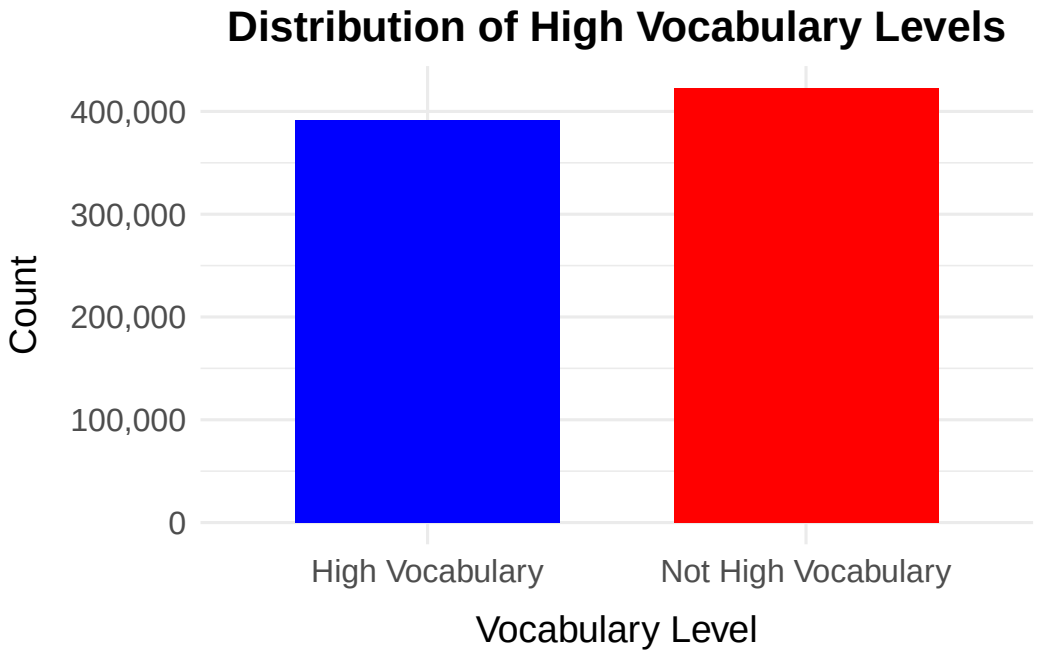

Figure 1: Distribution of the outcome variable, showing the counts of children classified as having "High Vocabulary" and "Not High Vocabulary" based on their comprehension and production scores. The bar plot illustrates the balance between the two categories in the dataset, which is important for modeling purposes.

2.4 Predictor variables

2.4.1 Age

Figure 2 displays the distribution of children's ages (in months) within the dataset, highlighting key patterns in the sample's demographic structure. A notable concentration of data is observed among children aged between 24 and 30 months, reflecting an emphasis on capturing

vocabulary development during critical periods of language acquisition. These age ranges are known to mark significant milestones in linguistic growth, which could explain their higher representation. Conversely, younger age groups (below 20 months) are underrepresented, likely due to the challenges of assessing vocabulary at earlier stages of development, where verbal communication is less pronounced and parental reporting is more variable.

The dataset also shows distinct peaks at specific ages, such as 25 and 30 months. These sharp spikes may reflect intentional focus points for testing or developmental benchmarks tied to standardized assessments like the MacArthur-Bates Communicative Development Inventories (CDI). This uneven age distribution underscores the importance of age as a critical factor in analyzing vocabulary acquisition. While the high concentration of data at older ages enhances insights into advanced vocabulary development, it also necessitates caution in generalizing findings to underrepresented age groups. This observation emphasizes the need to standardize age in statistical models to account for variability across different age groups.

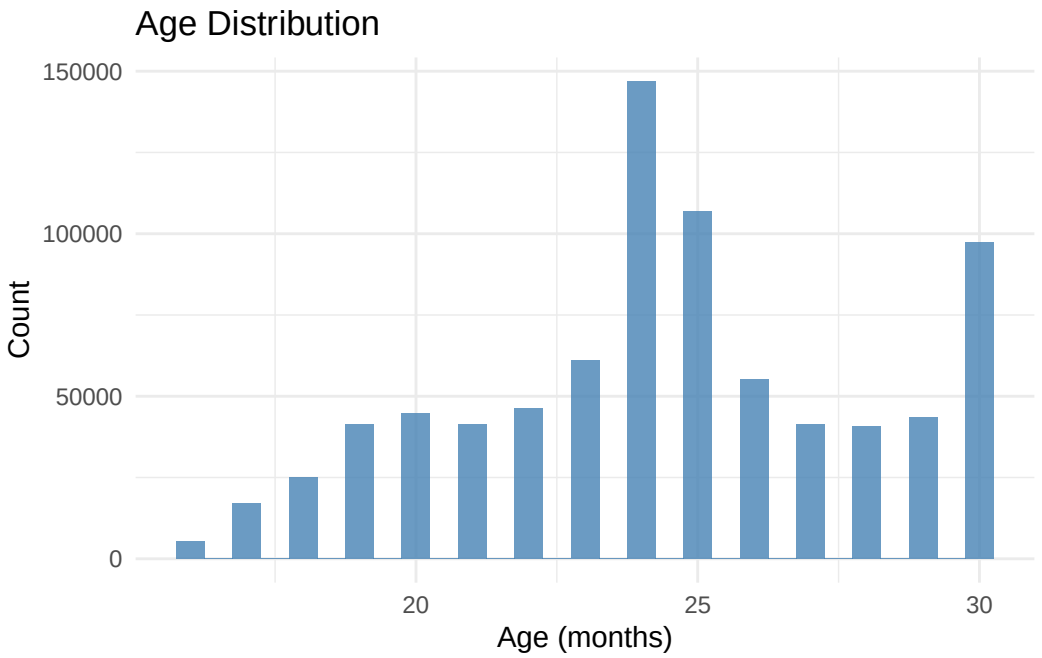

Figure 2: It shows the distribution of children's ages (in months) within the dataset. The majority of observations fall between 24 and 30 months, with noticeable peaks at 25 and 30 months, reflecting a focus on key developmental periods. Younger age groups are underrepresented, highlighting the need to account for variability in age when analyzing vocabulary acquisition patterns.

2.4.2 Broad Category

The words in the dataset were grouped into broad lexical categories to facilitate the analysis of vocabulary acquisition patterns. These categories include Activities, Adjectives, Function Words, Living Things, Objects, Places, Sensory Words, and Verbs. The classification was based on the semantic and functional roles of words, with nouns subdivided into more specific groups such as Living Things, Objects, and Places to capture distinct trends in vocabulary acquisition. For instance, Function Words include pronouns and question words, reflecting grammatical development, while Verbs and Adjectives capture action and descriptive words, essential for sentence construction and expression.

The bar graph illustrates the distribution of items across these categories, highlighting significant variation in word frequency. Objects constitute the largest category, suggesting a focus on tangible and concrete items, which are likely easier for children to recognize and recall. This is followed by Verbs and Living Things, categories that are fundamental to communication but slightly less prevalent. In contrast, Sensory Words and Activities are sparsely represented, possibly reflecting their specialized and context-dependent nature. The distribution underscores the importance of concrete and functional words in early vocabulary development while highlighting potential gaps in underrepresented categories. This variation provides a foundation for exploring how lexical diversity influences vocabulary acquisition patterns.

2.4.3 Norming Status

Norming status categorizes children into two groups: those included in the norming group and those who are not. The norming group represents a standardized sample used as a benchmark for assessing vocabulary development, providing a reference point for evaluating other children in the dataset. This distinction is important for ensuring the validity and reliability of comparisons in the analysis.

The bar plot illustrates the distribution of children based on their norming status. The dataset is predominantly composed of non-norming children, with only a small fraction representing the norming group. This imbalance reflects the practical challenges of including a broad and representative norming sample in large-scale vocabulary assessments. While the non-norming group provides diverse data for analysis, the norming group offers a crucial baseline for calibrating and interpreting the results. The distribution highlights the need to account for this imbalance in the analysis to avoid potential biases and ensure robust conclusions.

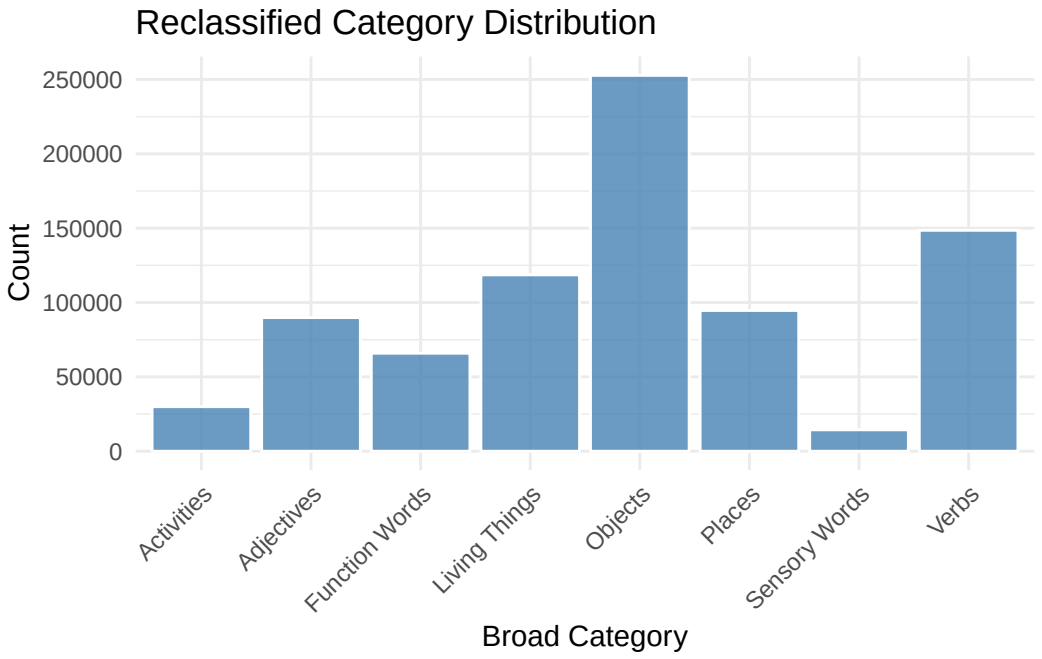

Figure 3: The figure shows objects dominate the vocabulary, reflecting an emphasis on concrete and tangible terms, while categories like Sensory Words and Activities are less frequently represented, indicating the relative complexity or specificity of these word types in early language acquisition

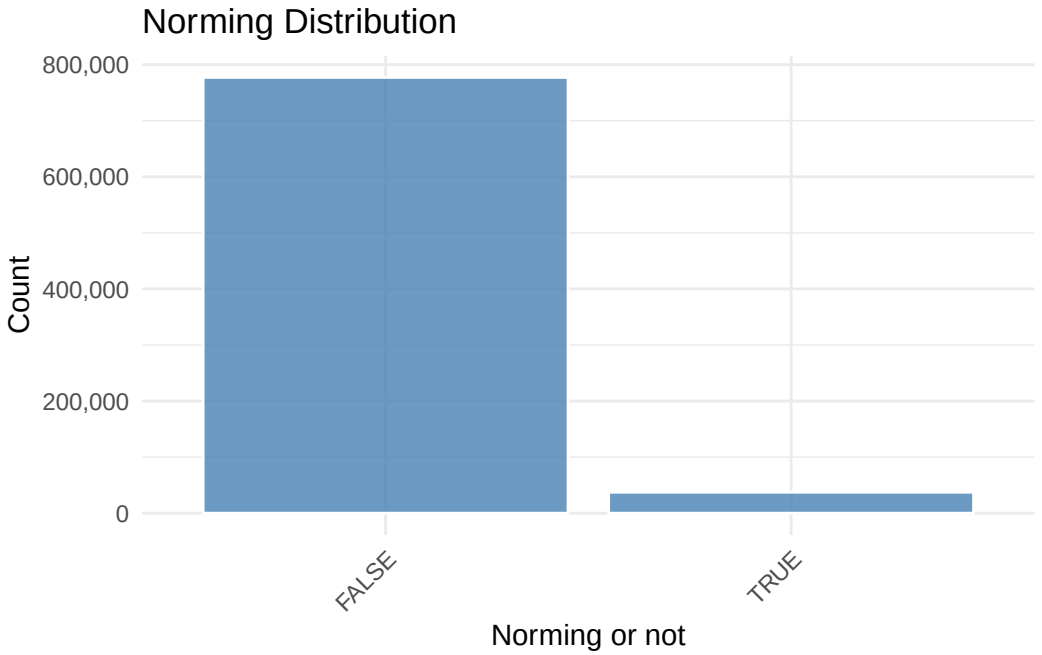

Figure 4: The dataset is primarily composed of non-norming children, with a smaller subset belonging to the norming group, serving as a standardized benchmark for assessing vocabulary development

3 Model

3.1 Model Selection

To investigate the relationship between children's vocabulary acquisition and their demographic and linguistic characteristics, we constructed a logistic regression model. 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 350. This threshold was chosen to distinguish children with relatively advanced vocabulary levels. More background details and diagnostics are included in Appendix- B.

3.2 Logistic Regression Model Overview

- High Vocabulary: The outcome variable, high_vocabulary, is a binary indicator that takes the value of 1 if the average production and comprehension score (prod_comp_mean) exceeds 350 and 0 otherwise. This threshold was selected to represent children with relatively advanced vocabulary skills, determined through exploratory data analysis.
- Scaled Age (age_scaled): This continuous variable represents the child's age, standardized to ensure the model coefficients reflect changes per standard deviation in age. Standardization improves numerical stability and aids interpretability.
- 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 four broad linguistic categories: nouns, verbs, adjectives, and function_words. The reference category is function_words.

The model takes the form:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \cdot \text{age\_scaled}_i + \beta_2 \cdot \text{is\_normingTRUE}_i \tag{1}$$

$$+ \beta_3 \cdot \text{broad\_categoryAdjectives}_i \tag{2}$$

$$+ \beta_4 \cdot \text{broad\_categoryFunction\_Words}_i \tag{3}$$

$$+ \beta_5 \cdot \text{broad\_categoryLiving\_Things}_i \tag{4}$$

$$+ \beta_6 \cdot \text{broad\_categoryObjects}_i \tag{5}$$

$$+ \beta_7 \cdot \text{broad\_categoryPlaces}_i \tag{6}$$

$$+ \beta_8 \cdot \text{broad\_categorySensory\_Words}_i \tag{7}$$

$$+ \beta_9 \cdot \text{broad\_categoryVerbs}_i \tag{8}$$

Where: - p_i represents the probability that person i has a high vocabulary - 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 - β_3, β_4, β_5, etc.: The effects of being in the respective broad word categories (nouns, function words, or verbs), compared to the reference category (likely "adjectives").

3.3 Model Assumptions

- Linearity of the Logit: The model assumes a linear relationship between the log-odds of the outcome (high vocabulary) and the independent variables. For example, the standardized age variable (age_scaled) assumes that for every one standard deviation increase in age, the log-odds of achieving a high vocabulary score increase by a constant amount. Standardizing age ensures that the variable is centered and scaled, making it easier to meet this linearity assumption and interpret its effect across the dataset.
- Independent Observations: The model assumes that all data points are independent. This assumption holds because each observation represents data from a unique child, with no repeated measurements for the same individual. For instance, there are no longitudinal observations or nested data structures (e.g., children grouped by classrooms or schools) that could violate independence. If dependence were present, a more complex model like a mixed-effects logistic regression would be necessary.
- Categorical Variable Encoding: The broad_category variable, which includes categories such as "Adjectives," "Verbs," and "Living Things," was encoded using sum contrasts. This approach ensures that the coefficients for each category represent the deviation of that category's effect from the overall mean effect across all categories. For example, the coefficient for "Verbs" indicates how the log odds of achieving a high vocabulary differ for "Verbs" compared to the average effect of all other categories. Sum contrasts are particularly useful for understanding relative effects and ensure that the intercept reflects the overall mean effect when all predictors are at their reference or average levels.

3.4 Interpretation of Coefficients

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

- Intercept (β_0): Represents the log-odds of high vocabulary when all predictors are at their reference or mean levels. If $\beta_0 > 0$, the baseline odds of high vocabulary are greater 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): If a child belongs to the norming group, the log-odds of high vocabulary increase by β_2 compared to non-norming children. If $\beta_2 = 0.1$, then $\exp(0.1) \approx 1.11$, meaning being in the norming group increases the odds of high vocabulary by 11%.
- Broad Category (β_3, β_4, β_5, etc.): The coefficients for `broad_category` represent the difference in log-odds compared to the reference category ("Adjectives").
  - β_3 (Function Words): A positive β_3 indicates higher odds of having a high vocabulary for function words compared to adjectives. For instance, if $\beta_3 = 0.002$, then $\exp(0.002) \approx 1.002$, meaning the odds of having a high vocabulary for function words are 0.2% higher than for adjectives. General Example for Broad Categories: If a coefficient $\beta_k = 0.01$, $\exp(0.01) \approx 1.01$, meaning the corresponding category increases the odds of having a high vocabulary by 1% compared to the reference category (Adjectives). Conversely, if $\beta_k = -0.01$, $\exp(-0.01) \approx 0.99$, indicating a 1% decrease in odds compared to the reference category.

3.5 Model Justification

The model was trained on 80% of the dataset, with the remaining 20% reserved for testing. Model performance was assessed using confusion matrices and overall accuracy. Further evaluation included the analysis of residual deviance, AIC, and interpretation of individual coefficients.

4 Results

4.1 Average Production by Broad Category and Age

Figure 5 shows..

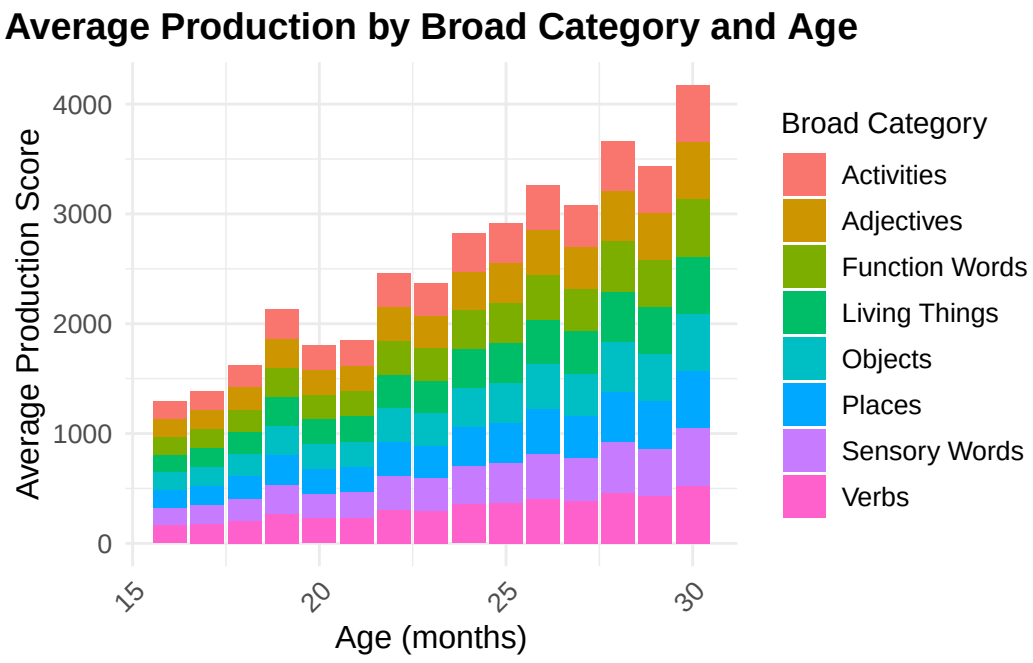

Figure 5: xx

4.2 Prediciton for the Probability of High Vocabulary Level

```
      Predicted
Actual     0     1
     0 59837 25062
     1 26778 51251
```

```
[1] "Accuracy: 0.68182264558578"
```

4.3 Distribution of Predicted Probabilities by Broad Category

The Figure 6 illustrates the distribution of predicted probabilities for high vocabulary levels across various broad categories. The density curves show distinct peaks and variability, reflecting the differences in how well each category predicts high vocabulary. Notably, "Sensory Words" and "Objects" exhibit higher density in the middle range of predicted probabilities, suggesting moderate association with high vocabulary. In contrast, categories like "Activities" and "Function Words" have wider, flatter distributions, indicating greater uncertainty or diversity in prediction. This variability highlights the nuanced role of word categories in predicting high vocabulary acquisition. Further analysis could explore why certain categories contribute more consistently to predictions than others. The density plot highlights distinct

patterns in predicted probabilities, with nouns and verbs showing higher peaks, indicating a stronger likelihood of high vocabulary acquisition in these categories. This suggests that broad categories contribute differently to vocabulary development, with nouns and verbs potentially playing a more significant role.

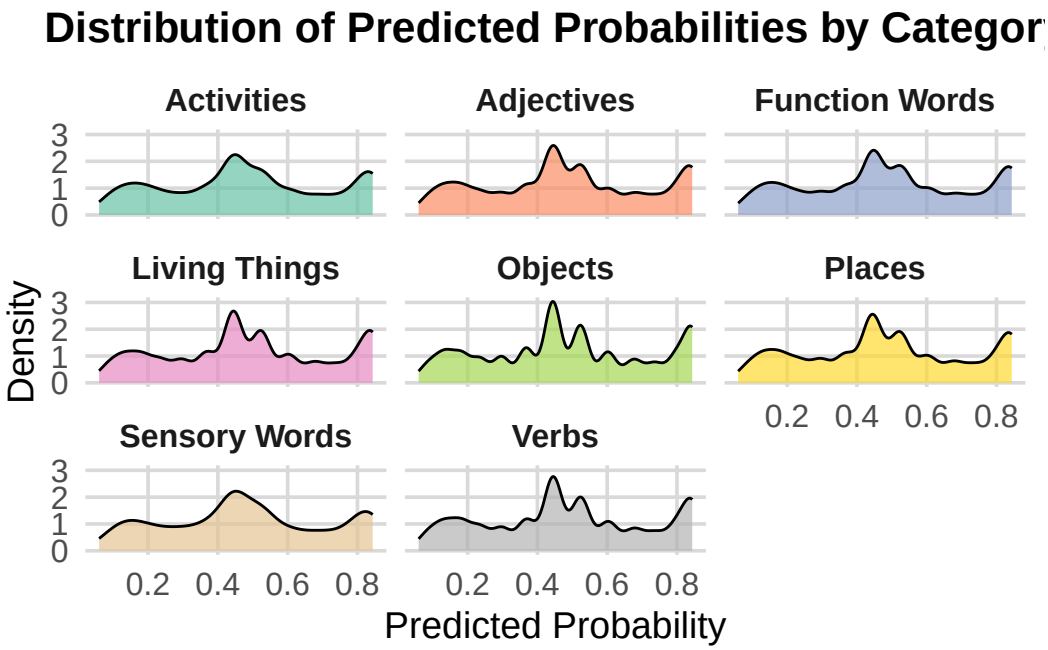

Figure 6: This density plot illustrates the predicted probabilities of having a high vocabulary across different broad lexical categories. Each curve represents the density of predicted probabilities within a category, showcasing the variation in predicted outcomes for categories such as Activities, Adjectives, Function Words, Living Things, Objects, Places, Sensory Words, and Verbs. The plot highlights overlapping patterns and areas of divergence in vocabulary acquisition likelihood across different types of words.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

A Appendix

A.1 Additional data details

A.2 Data Sheet

B Model details

B.1 Model Summary

B.2 Diagnostics

Table 2: model summary table

	(1)
(Intercept)	−0.092
	(0.015)
age_scaled	1.124
	(0.003)
is_normingTRUE	−0.168
	(0.013)
broad_categoryAdjectives	−0.004
	(0.017)
broad_categoryFunction Words	0.002
	(0.018)
broad_categoryLiving Things	−0.008
	(0.016)
broad_categoryObjects	−0.008
	(0.015)
broad_categoryPlaces	−0.009
	(0.017)
broad_categorySensory Words	−0.004
	(0.026)
broad_categoryVerbs	−0.005
	(0.016)
Num.Obs.	651 712
AIC	752 204.6
BIC	752 318.5
Log.Lik.	−376 092.312
RMSE	0.44

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), (**modelsummary?**) 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.

A special acknowledgment goes to the Braginsky (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

Braginsky, Mika. 2024. *wordbankr: Accessing the Wordbank Database*. https://github.com/langcog/wordbankr.

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, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

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.

Xie, Yihui. 2024. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.