Identifying the distributional sources of children's early vocabulary

Anonymous CogSci submission

Abstract

Children's early word learning is to a large extent driven by the prevalence of words in their language environment, with words that are spoken more often to children being learned earlier. However, children receive language from a variety of sources, including books, television, and movies meant for children, as well as speech and media that is meant for adults, but overheard by children. Despite considerable similarity of word frequency distributions from these different input sources, there is also significant and predictable variability between them. For example, function words are far more frequent in books than in everyday speech, while early-learned nouns (e.g., 'ball' and 'mommy') are more frequent in child-directed speech than from other sources. Children receive a mixture of these different frequency distributions, and the ratios of the mixture may be predictive of their early word learning. The goal of this paper is to better understand the shared and unique variance in these sources of input–in both English and French–and to evaluate how predictive these input frequencies are of children's early word learning.

Keywords: early language learning; CDI; vocabulary development; word frequency distributions.

Introduction

How does speech addressed to children, heard on television, or read in books impact the growth of children's early vocabulary? How does speech from these sources relate to adult-directed sources of speech? And how do these potential language sources combine with parental education to predict young children's vocabulary growth? Children must learn words based on ambient linguistic input, and indeed the amount of child-directed speech a child receives predicts later vocabulary growth (Hart & Risley, 1995). However, children's exposure to different words can vary greatly depending on how much child-directed speech they hear, how much they are read to, and how much they are overhearing adults speaking to each other. These different language input sources can vary both in quantity, as well as in various measures of quality (e.g., diversity of the types of words used, the length of utterances), which may influence children's uptake. Moreover, the amount of input children receive from these different input sources may vary from child to child, which may account for some of the great variability seen in children's early vocabulary growth (cf. Fenson et al., 1994). Indeed, higher measures of input quantity and quality have been found to relate to children's faster vocabulary growth, and to often be related to parents' socioeconomic status (SES; Rowe, 2012; Hoff, 2003).

Input word frequency varies significantly depending on the context. Previous studies have shown that frequency matters for children's word learning (for a review, see Ambridge, Kidd, Rowland, & Theakston, 2015), and have observed an association between word frequency in children's language environments and age of acquisition (Goodman, Dale, & Li, 2008). For instance, word frequency in books is not the same as frequency in conversational speech, with many function words being far more frequent in books than in speech (Dawson, Hsiao, Wei Ming Tan, Banerji, & Nation, 2021; Montag, Jones, & Smith, 2015).

Some differences between frequency distributions are intuitive: "mommy" is quite frequent in child-directed speech, yet not so common in children's books, and even more rare in books meant for all ages. But other differences are less intuitive: "of" is frequent in books meant for all ages, and while still frequent in child-directed speech, it is relatively less frequent as compared to children's books. In general, speech–whether directed to children, or to adults–contains relatively fewer function words, and tends to score lower on measures of lexical diversity than books, which have a higher ratio of types (i.e. unique words) per set of tokens (i.e., instances of words; Dawson et al., 2021).

The primary questions of this paper are, first, to examine shared and unique variance in word frequency across different sources of English and French input, ranging from children's books and movies to child-directed speech and even comparing to adult-directed books, movies, and speech, which we accomplish using principle components analysis (PCA).

Second, we investigate how well these components predict English- and French-learning children's early word learning, using aggregate MacArthur-Bates Communicative Development Inventories (CDI) data from Wordbank (Frank, Braginsky, Yurovsky, & Marchman, 2017). The CDIs have proven to be reliable and valid indicators of child's language, with high internal consistency and predictive of later language outcomes (Fenson et al., 1994). We approximate early word learning with the Age of Acquisition (AoA) prediction paradigm, which consists in predicting each CDI item's mean Age of Acquisition (AoA) – the mean age (in months) at which 50% of children are expected to know a given word (Braginsky, Yurovsky, Marchman, & Frank, 2019; Goodman et al., 2008).

Third, we examine how well the frequency components

predict individual differences in English-learning children's word learning in combination with their mother's education, which may be related to how much children are read to at home. Past research has found that young children from higher-SES households tend to have larger vocabulary (Fernald, Marchman, & Weisleder, 2013), and parents with higher-SES tend to report reading more to their young children than parents with lower SES. Can vocabulary composition of children from higher-SES households be better predicted by book word frequency?

The questions investigated in this paper aim to shed light on the importance of input sources on language development and its interaction with parental SES, would can be useful for informing future interventions.

Method

Datasets

Corpora from different sources are used to identify shared and distinct variance in frequencies, with the help of PCA.

Child-directed Speech (ChS). Utterances of ChS were extracted from the CHILDES corpus (MacWhinney, 2000), a collection of transcripts of interactions between caregivers and children of ages ranging from 0 to 12 years ($M = 2.9$ years). After cleaning, the CHILDES English corpus yielded 5521000 tokens across 38779 word types. The French ChS yielded 1000000 tokens across 14353 word types.

Child-directed books (ChB). We used a sample of 98 English children's books from Project Gutenberg's open-source database, previously used in machine learning research on language comprehension (Hill, Bordes, Chopra, & Weston, 2015). The books were published between 1820 and 1922, but include well-known titles as *The Legend of Sleep Hollow*. We also used 130 popular French children stories accessible in parenting websites (`https://fr.hellokids.com/`) and 10 French children books from Project Gutenberg. After cleaning, the English ChB corpus totals 4674000 tokens across 42666 word types, and the French ChB totals 1000000 tokens across 17852 word types.

Child-directed Media (ChM). Transcripts were extracted from English television shows (e.g., from PBS Kids and Nickelodeon) and movies (e.g., *Beauty and the Beast*), including 1,078 movies and 4,309 TV episodes taken from Charlesworth, Yang, Mann, Kurdi, & Banaji (2021) (available here: `https://osf.io/kqux5/`. Openly accessible transcripts (`https://www.subsynchro.com/`) were also extracted from 100 French films directed to children. After cleaning, the English ChM totals 6759247 tokens across 84333 word types. The French ChM totals 1000000 tokens across 14843 word types.

Adult-directed Speech (AdS). English AdS was obtained from the Switchboard-1 Telephone Speech Corpus (Godfrey & Holliman, 1993), a corpus of transcripts from dyadic telephone conversations. French AdS was obtained from the TCOF corpus (André & Canut, 2010), the CLAPI corpus (Balthasar & Bert, 2005) and the CFPP corpus (Branca-Rosoff, Fleury, Lefeuvre, & Pires, 2012). After cleaning, the English AdS yielded 3104000 tokens across 27536 word types. The French AdS yielded 1000000 tokens across 14391 word types.

Adult-directed Books (AdB). The English AdB is comprised of 35144000 tokens across 827414 word types. The French AdB is comprised of books taken from the 1999 Association de Bibliophiles Universels, an open-source database of french books. After cleaning, it yielded 1000000 tokens across 30447 word types.

Adult-directed Media (AdM). The English AdM is comprised of 6167000 tokens across 62876 word types. The French AdM corpus is comprised of 1000000 tokens across 17843 word types, after cleaning openly accessible movie subtitles (`https://www.subsynchro.com/`) from 100 films.

Merging the Corpora

Children's early word learning data is drawn from the CDIs (Fenson et al., 2007), aggregated in the Wordbank database (Frank et al., 2017) (data from X kids for American English and X kids for French French). CDIs are parental reports on their children's lexical development, proven to be reliable indicators of a child's language. CDIs are survey instruments, where parents mark whether their child (age ranges 8-15 and 16-30 months old) understands or produces particular words out of a list of several hundred words. All word frequencies were normalized to number of tokens per million (TPM). We focus our analysis on the 674 words from the English CDI that we were able to find in at least some of the corpora, and 470 words from the French CDI (for French, words were matched to related words in corpora via a stemmer). For any CDI words that failed to appear in a given corpus, we replaced the missing word's frequency with a normalized count of 10 TPM, or the minimum normalized frequency for that distribution, whichever was smaller.

Results

Cross-corpus Frequency Correlations (Q1)

Figure 1 shows the word frequency correlations between different corpus sources ([Adult- vs. Child-directed] x [Speech, Books, Media]) for the matched CDI words (left: English, right: French).

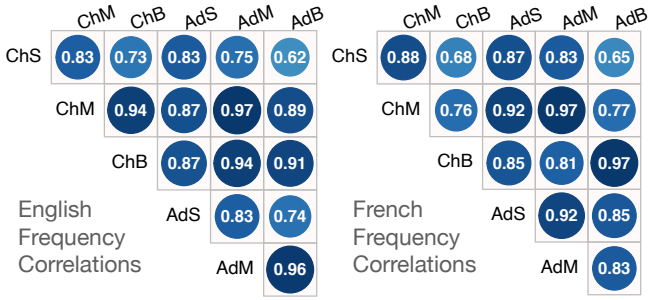

Figure 1: Word frequency correlations between different corpus sources for the matched CDI words in English (left) and French (right).

For both languages, there are both register and source effects. Unsurprisingly, there were strong correlations across these different corpora. We thus turned to PCA to disentangle these correlated distributions and to understand their interrelations.

Table 1 shows the standard deviation (Std Dev) and proportion of variance explained, both individually (Prop Var) and cumulatively (Cum Prop Var) by the principal components (PC1-PC6) for both languages. PC1 already explains the bulk the variance (89% for English and for French), and PC2-PC4 each only capture an additional 2-4% of the variance. In total, the first four components captured >98% of the variance for both languages.

Table 2 shows the eigenvectors of the principal components (PC1-PC6) in relation to the original six frequency distributions for English and French The first PC is similar for both English and French, and captures shared variance between all frequency sources, representing words that are high or low frequency across registers and sources. This component captures for example that 'the' in English or 'le' ou 'et' = and in French is frequent in spoken and written text and in speech to adults and children. For English, PC2 (3.8% of variance) mostly captures child-directed speech, differentiating it from the other registers. For example, words such as 'peekaboo' are mostly present in this speech to children. For French, PC2 (8% of variance) mostly captures books for children and adults. This component captures that some words are very frequent in text in general, such as 'sombre' = 'dark', 'lune' = 'moon' or 'autrement' = 'otherwise' comparing to the other registers.

For English, PC3 (3% of variance) captures adult-directed speech, differentiating it from the other registers. It captures words such as 'mower', 'downtown' or 'vitamins', frequent in discussions between adults. For French, PC3 (4%) is similar to PC2 in English, capturing child-directed speech, with words such as 'gâteau' = 'sweet'. For English, PC4 (2.2% of variance) captures the similarity of child- and adult-directed media, distinguishing them from child- and adult-directed books. This component captures that some words are very frequent in media in general, but infrequent in books (e.g. 'camera', 'tv') and vice versa (e.g. 'none', 'beads', 'poor'). For French, PC4 (4% of variance) captures the similarity of child- and adult-directed media (e.g. 'étoile' = 'star'), distinguishing them from child- and adult-directed speech (e.g. 'voiture' = 'car').

For English, PC5 (1% of variance) captures child-directed speech, distinguishing it from child-directed books. his component captures that some words are very frequent in child-directed speech, but infrequent in child-directed books (e.g.'mommy', 'juice') and vice versa (e.g. 'firetruck', 'snowsuit'). For French, PC5 (3% of variance) distinguishes child-directed media (e.g. 'ferme' = 'farm') from adult-media (e.g. 'café). Last, for English, PC6 (<1% of variance) is similar to PC5 in French, capturing child-directed media (.73), and distinguishing it in particular from adult-directed media. This component captures that some words are very frequent in child-directed media, but infrequent in adult-directed media (e.g. 'penguin', 'snowman') and vice versa (e.g.'drawer', 'medicine'). For French, PC6 (3%) distinguishes child-directed books (e.g. 'ours' = 'bear') from adult-directed books (e.g. 'tuer' = 'kill').

In summary, the PCs align surprisingly well with particular dimensions of the English frequency distributions: PC1 with overall frequency, PC2 with child-directed speech, PC3 with adult-directed speech, PC4 with media vs. books, PC5 with child-directed books vs. speech, and PC6 with adult- vs. child-directed media. For French, we observe a similar pattern of findings, although the importance and order of the PCs is not the same. Variance is distributed across the PCs, frequency differences across different registers being more pronounced than for English. Importantly, frequency particularities in ChS come third in order in French, compared to second in English. On the contrary, word frequency in books differs from all other registers in French, and comes second in order.

Table 1: Importance of components from PCA for English (above) and French (below).

	PC1	PC2	PC3	PC4	PC5	PC6
StdDev Eng	5.10	1.05	0.97	0.82	0.57	0.44
PropVar Eng	0.89	0.04	0.03	0.02	0.01	0.01
CumPropVar Eng	0.89	0.93	0.96	0.98	0.99	1.00
StdDev Fr	2.87	0.75	0.53	0.41	0.30	0.25
PropVar Fr	0.88	0.06	0.03	0.02	0.01	0.01
CumPropVar Fr	0.88	0.94	0.97	0.98	0.99	1.00

PCA-based Age of Acquisition Regression (Q2)

Next, we turned to question 2 asking how well these components predict English- and French-learning children's early word learning. We attempted a simple regression predicting each CDI word's mean AoA. However, multicollinearity makes it unwise to include these raw distributions in a regression, as the results would likely be unstable. We verified

Table 2: Highlights from principal components' rotation in the original coordinate system.

Order	EN	FR
PC1	Freq -.26...-.52	Freq -.38...-.42
PC2	ChS .65	B -.52,-.56
PC3	AdS .79	ChS .65
PC4	M/B .40.48/-.44-.63	M/S .65.23/-.20-.68
PC5	ChS/ChB -.60/.53	ChM/AdM -.53/.69
PC6	ChM/AdM .73/-.53	ChB/AdB -.59/.70

that this was the case by running a regression predicting AoA with log(word frequency) from each of the six distributions as predictors.

The Variance Inflation Factor (VIF) estimates the inflation of the variance of a regression coefficient when there is correlation between predictors (Dodge, 2008). The more the VIF increases, the less reliable the regression results are. The VIF for every distribution was $>> 1$ (and many > 5 or > 10), indicating that these variables show strong multicollinearity which may compromise the reliability of the regression results. We thus used the CDI items' PCA loadings in lieu of the frequency distributions to predict AoA. Figures showing the loadings of CDI items on PC1-PC4 vs. the average age of acquisition (AoA; in months) can be found in the appendix (link).

Given that past research has found that lexical class strongly modulates influences of word frequency, we next examined the interaction of lexical class (LC) with PC1 - PC6 in our regression. We also included the number of letters as a predictor (Nletters) to help control for the overall difficulty of each word. To determine if the inclusion of all PCs was justified, we ran a series of ANOVAs building up from PC1 to PC6–in decreasing order of the variance they accounted for in the PCA[1]. The more complex model was always significantly preferred, including up to the inclusion of PC6 (English: $R^2 = .584$, French: $R^2 = .16$). Figure 2 shows the coefficient estimates with $p<0.05$ for both languages.

[1]The R syntax for the sequence of regressions was `AoA~PC1*LC`, `AoA~(PC1+PC2)*LC`, ..., `AoA~(PC1+PC2+PC3+PC4+PC5+PC6)*LC`, with noun as the baseline LC.

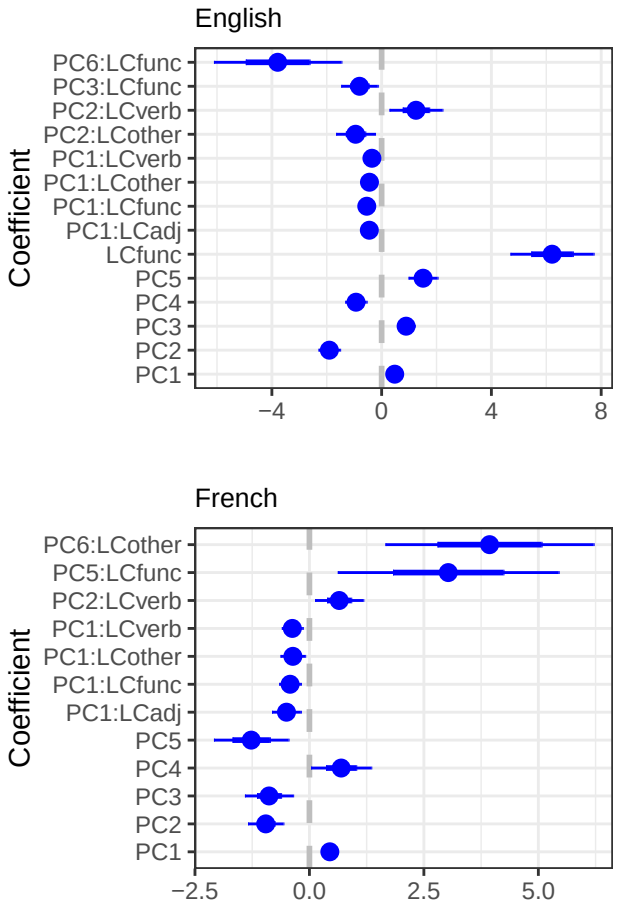

Figure 2: Regression coefficients for predicting CDI AoAs with PCs and lexical class for English (top) and French (bottom).

In general, we observe that registers and sources both matter for early word learning. For English, P1, P2, P3, P4 and PC5 predict the age of acquisition. Overall frequency (PC1) is a predictor; words which are frequent in general are learned earlier than less frequent words. Child-directed speech (PC2) is a negative predictor; words which are frequent in this register are learned earlier on. On the contrary, adult-directed speech is a positive predictor; words which are frequent in this register are learned later on. The frequency of words in media distinguished from books is a predictor (PC4). Frequency of words in child-directed speech distinguished from words in child-directed books is also a significant predictor; an earlier acquisition is predicted for words not existing in books. We also observe that overall frequency interacts with lexical class, verbs, function words and adjectives being learned later than nouns. PC2 interacts with verbs, which are learned later than nouns. There is also an interaction with PC3 and function words, which are learned later than nouns, and PC6 and function words.

For French, registers as well as sources in are also important predictors of word learning. P1,P2 P3 and PC4 and PC5 are significant predictors, which means that speech, books, media and the child-directed register are important. Overall

frequency (PC1) is a predictor; words which are frequent in general are learned earlier than less frequent words. Book frequency (PC2) is a predictor; words which are especially frequent in books are learned later on. Child-directed speech (PC3) is a predictor; words which are especially frequent in this register are learned earlier on. PC4 includes words which are often found in media versus speech, and PC5 includes words in child-directed versus adult-directed media. Similarly to English, we also observe that overall frequency interacts with lexical class, verbs, function words and adjectives being learned later than nouns. PC2 interacts with verbs, which are learned later than nouns. PC5 also interacts with function words. There is no significant effect of word length.

In sum, in both languages, we observe that the predictive value of word frequency comes from speech, media and book sources, and mostly the child-directed register. The main differences across languages are that for English, the difference and difficulty of adult-directed speech is emphasized. For French, the difference and difficulty of text, independent of register, is emphasized.

Combining distributions with demographic data (Q3)

Previous findings relate SES status to book reading (Shen & Del Tufo, 2022), which is in turn related to better language skills (Bus, Van Ijzendoorn, & Pellegrini, 1995). This suggests that the vocabulary composition of children from higher-SES households may be better predicted by the word frequencies seen in child-directed books, rather than those from child-directed speech. To test this idea, we did an exploratory logistic regression using the first four PCs to predict the number of children in Wordbank who produce or don't produce each item, along with interactions of mother's education and children's age. Due to lack of maternal education data for French data, we focus on American English-learning children. American English data from Wordbank contained 2,776 CDI:WS administrations with mother's education coded as a factor (baseline (0): no more than secondary education (N=547), 1 for some/all college (N=1483), and 2 for at least some graduate school (N=746)).

There were significant main effects of age, mother's education, and PC1-PC4 (all $p < .001$). There were significant interactions of age with mother's education ($p < .001$), PC2 ($p < .001$), and PC4 ($p = .009$). There were significant interactions of mother's education with PC1-PC3 (all $p < .001$), shown in Figure 4.

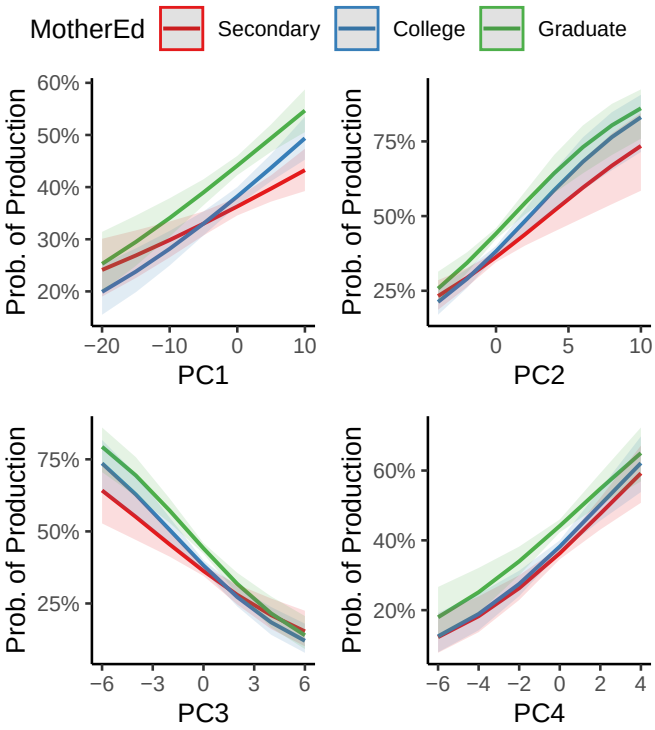

Figure 3: Predicted effects of maternal education and PC1-PC4 on the probability of children producing CDI words.

Discussion

We set out to investigate the sources of linguistic input and registers that children may experience, using word frequency distributions garnered from child-directed and adult-directed corpora of speech, books, and media (TV and movies). We found the principal components (PCs) of these distributions, and described how these PCs capture variation both in adult- vs. child-directedness, as well as between modalities (e.g., books and speech).

Our findings show that multiple components are predictive of children's age of acquisition of words from the CDI. Speech, media and books are all relevant in predicting age of acquisition, as well as child-directed speech, which suggests that children's environmental exposure to different language source could impact their word learning trajectory.

We also used English Wordbank data to examine how well the frequency components combine with mother's education–a measure of SES that has been found in the past to be positively related to early word learning, and to more child-directed reading–to predict children's early word learning. This analysis revealed significant contributions of multiple PCs, as well as the interaction of mother's education with PC1 (overall frequency), PC2 (child-directed speech), and PC3 (adult-directed speech), but not of the child-directed books component (PC4).

These findings are important for educational reasons and could inform future interventions on the importance of including language from different sources. A target for future

research is to predict individual children's learning of particular words using these principal components, in combination with parent-reported measures of how much time their child spends daily receiving input from each of the input sources ([adult- vs. child-directed] x [books, media, and speech]).

Finally, there was overwhelming similarity of word frequency distributions from different sources and their predictive value in English and French data, which supports the robustness of the results. At the same time, these distributions show systematic variation–at least in English and in French, which can predict some of the variation in children's early word learning. We should also notice that different data and corpus sizes were used to represent the sources for each language. Differences could also be partly attributed to this e.g., the French corpora were composed of several smaller ones, due to the lack of large accessible corpora, and had slightly different use e.g. small stories in French, large books in English.

By better understanding the similarity and differences between word frequencies children experience in different contexts, future research in this vein holds the promise to predict individual differences in children's early word learning on the basis of their daily routines.

Acknowledgements

[Redacted for anonymous review.]

References

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*(2), 239–273.

André, V., & Canut, E. (2010). Mise à disposition de corpus oraux interactifs: Le projet tcof (traitement de corpus oraux en français). *Pratiques. Linguistique, Littérature, Didactique*, (147-148), 35–51.

Balthasar, L., & Bert, M. (2005). La plateforme corpus de langues parlées en interaction(CLAPI). Historique, état des lieux, perspectives. *Lidil. Revue de Linguistique et de Didactique Des Langues*, (31), 13–33.

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, *3*, 52–67.

Branca-Rosoff, S., Fleury, S., Lefeuvre, F., & Pires, M. (2012). Discours sur la ville. Présentation du corpus de français parlé parisien des années 2000 (cfpp2000). *Article En Ligne, Http://Cfpp2000. Univparis3. Fr/Articles. Html*.

Bus, A. G., Van Ijzendoorn, M. H., & Pellegrini, A. D. (1995). Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Review of Educational Research*, *65*(1), 1–21.

Charlesworth, T. E., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science*, *32*(2), 218–240.

Dawson, N., Hsiao, Y., Wei Ming Tan, A., Banerji, N., & Nation, K. (2021). Features of lexical richness in children's books: Comparisons with child-directed speech. *Language Development Research*.

Dodge, Y. (2008). *The concise encyclopedia of statistics*. Springer Science & Business Media.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, i–185.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: User's guide and technical manual (2nd ed.)*. Baltimore, MD: Brookes.

Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, *16*(2), 234–248.

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

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

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young american children.* Paul H Brookes Publishing.

Hill, F., Bordes, A., Chopra, S., & Weston, J. (2015). The goldilocks principle: Reading children's books with explicit memory representations. *arXiv Preprint arXiv:1511.02301*.

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*(5), 1368–1378.

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk. Transcription format and programs* (Vol. 1). Psychology Press.

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.

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development*, *83*(5), 1762–1774.

Shen, Y., & Del Tufo, S. N. (2022). Parent-child shared book reading mediates the impact of socioeconomic status on heritage language learners' emergent literacy. *Early Childhood Research Quarterly*, *59*, 254–264.