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. 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 the source – spoken language vs. books vs. media – and the register – child-directed vs. adult directed – of the language. 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 (unique words) per set of tokens (instances of words; Dawson et al., 2021).

In this paper, we have three goals. First, we 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. Because of the substantial correlations between these different input sources, we employ principal components analysis (PCA) for dimensionality reduction.

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). CDIs are parent report forms for children's early vocabulary, and they have proven to be reliable and valid indicators of child's language (Fenson et al., 1994). Critically, CDI forms provide details about the individual words that children produce. These data allow us to investigate the role of different frequency sources using the Age of Acquisition (AoA) prediction paradigm, in which we use regression models to predict 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 different input sources predict SES differences in word learning for English-learning

children. We index SES using maternal education, a common proxy measure. 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.

Method

Datasets

Corpora from different sources are used to identify shared and distinct variance in frequencies. These corpora vary widely in size due to data accessibility; several were created for the current study and are available in our GitHub repository.

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 3102000 tokens across 13016 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 Sleepy 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 4673000 tokens across 42444 word types, and the French ChB totals 1298000 tokens across 17990 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 6723850 tokens across 80082 word types. The French ChM totals 842000 tokens across 14937 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 27479 word types. The French AdS yielded 1466000 tokens across 14486 word types.

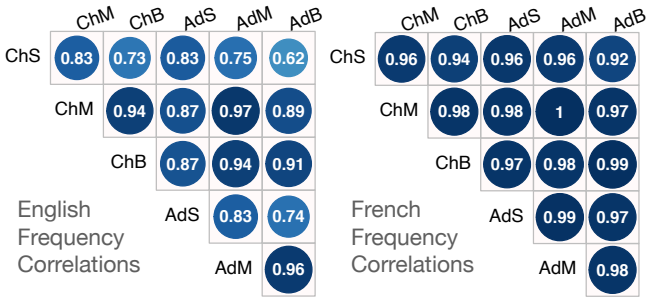

Figure 1: Word frequency correlations between different corpus sources for the matched CDI words in English (left) and French (right).

Adult-directed Books (AdB). The English AdB is comprised of 33800000 tokens across 147937 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 2288000 tokens across 30615 word types.

Adult-directed Media (AdM). The English AdM is comprised of 6060000 tokens across 60626 word types. The French AdM corpus is comprised of 766000 tokens across 15662 word types, after cleaning openly accessible movie subtitles (`https://www.subsynchro.com/`) from 100 films.

Age of Acquisition data 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 5520 children aged 16-30 months for the American English CDI: Words & Sentences (WS) form, and 641 children for the French French CDI:WS form). Age of acquisition estimates were calculated via the wordbankr package using the XYZ method.

Merging the Corpora

We focus our analysis on the 670 words from the English CDI that we were able to identify in at least some of the corpora, and 632 words from the French CDI. For French, because of the presence of more complex morphology, CDI 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. All word frequencies were normalized to number of tokens per million (TPM).

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). Unsurprisingly, there were strong correlations across these different corpora, but correlations were stronger

within register and within source for both English and French.

We thus turned to PCA to disentangle these correlated distributions and to understand their relationship. PCA allows us to project word frequencies into a space in which the first dimension captures the shared variance between frequencies from different sources and registers, and subsequent dimensions capture other consistent sources of variation. Since the logarithm of frequency is typically used as a psycholinguistic predictor in previous studies (Braginsky et al., 2019; Goodman et al., 2008), we perform our PCA over log frequencies.

Table 1 provides descriptions of the principal components (PC1-PC6) for each language. Table 2 summarizes the eigenvectors of the PCs in relation to the original six frequency distributions for English and French, and shows the proportion of variance explained by each. PC1 already explains the bulk of the variance (89% for English and 90.1% for French); and PC2-PC4 each only capture an additional 2-4% of the variance for both languages.

The first PC is similar for both English and French, and captures shared variance between all frequency sources and registers, representing words that are high or low frequency across them. This means that frequency distributions are largely similar across sources and registers. The second PC for English mostly captures child-directed speech, differentiating it from ll the other registers, whereas for French it captures book language, differentiating it from speech. PC2 explains 3.8% of variance for English and 4.4% of variance for French. Adult-directed speech is captured in English as the third PC (3.2% of variance), whereas the difference between child-directed and adult-directed speech is captured in French as the third PC (2.6% of variance). The additional PCs capture differences between media and book or speech, as well as differences between media registers (child-directed vs adult-directed) for both languages.

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. A difference lies on PC2; whereas it captures register differences, especially for child-directed speech comparing it to all other sources in English; it captures source differences in French, distinguishing book language from speech.

PCA-based Age of Acquisition Regression (Q2)

Next, we turned to our second question: how well different frequency distributions predict English- and French-learning children's early word learning. Our approach was to fit a linear regression model predicting each CDI word's mean AoA, following previous work (Braginsky et al., 2019; Goodman et al., 2008).

Multicollinearity makes it unwise to include multiple raw frequency distributions in a regression, however, as the results will be unstable. We verified that this situation was the

Table 1: Descriptions of English and French PCs

Order	Description	Example
PC1_EN	overall freq	the (EN)
PC1_FR	overall freq	le (FR the)
PC2_EN	CDS	peekaboo (EN)
PC2_FR	Books vs Speech	autrement vs gâteau (FR otherwise, sweet
PC3_EN	ADS	mower (EN)
PC3_FR	CDS vs ADS	xxxxx (FR)
PC4_EN	Media vs Books	camera vs beads (EN)
PC4_FR	Media vs Speech	étoile vs voiture (FR star, car)
PC5_EN	CDS vs CDB	juice vs firetruck (EN)
PC5_FR	CDM vs ADM	ferme vs café (FR farm, cafe)
PC6_EN	CDM vs ADM	snowman vs medicine (EN)
PC6_FR	CDB vs ADB	ours vs tuer (FR bear, kill)

Table 2: English principal component rotations.

	PC1	PC2	PC3	PC4	PC5	PC6
ChM	*-0.31*	0.15	**-0.34**	0.40	0.28	**0.73**
ChB	*-0.34*	0.25	-0.32	-0.63	**0.53**	-0.21
ChS	*-0.26*	**0.65**	-0.30	0.12	**-0.60**	-0.22
AdM	*-0.47*	-0.45	-0.24	**0.48**	0.13	**-0.53**
AdB	*-0.49*	**-0.47**	-0.01	-0.44	-0.50	0.32
AdS	*-0.52*	0.27	**0.79**	0.10	0.14	-0.01
PVar	0.89	0.04	0.03	0.02	0.01	0.01

Table 3: French principal component rotations.

	PC1	PC2	PC3	PC4	PC5	PC6
ChM	*-0.41*	-0.09	0.05	**-0.65**	**0.56**	0.31
ChB	*-0.41*	**0.54**	0.21	0.13	0.26	**-0.64**
ChS	*-0.36*	**-0.51**	**0.73**	0.22	-0.16	0.02
AdM	*-0.42*	-0.19	-0.34	-0.42	**-0.62**	-0.34
AdB	*-0.41*	**0.54**	0.02	0.18	-0.36	**0.62**
AdS	*-0.42*	-0.34	**-0.55**	**0.55**	0.30	0.04
PVar	0.90	0.04	0.03	0.02	0.01	0.00

case by running a regression predicting AoA with the logarithm of word frequency from each of the six distributions as a predictor. The Variance Inflation Factor (VIF) estimates the inflation of the variance of a regression coefficient when there is correlation between predictors (Dodge, 2008). The higher the VIF for a predictor, the less reliable the regression results are when that predictor is included. The VIF for every distribution was $>> 1$ (and many > 5), 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.

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]. For English, the more complex model was always significantly preferred, including up to the inclusion of PC6 ($R^2 = .58$). The same was observed for French, indicating that registers and sources both matter for early word learning. However, the french model explained less variation of the dependent value overall ($R^2 = .09$). Figure 2 shows the coefficient estimates with $p<0.05$ for both languages.

For English, PC1, PC2, PC3, PC4 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 predictor; words which are frequent in this register are learned earlier on. On the contrary, in the adult-directed speech predictor (PC3), frequent words are learned later on. Word frequency in media distinguished from words in books is a predictor (PC4); words which are frequent in media as distinguished by words in books, are learned earlier on. Word frequency in child-directed speech as distinguished from words in child-directed books is also a predictor, with earlier acquisition predicted for more speechy/less booky words. We also observe that overall frequency (PC1) interacts with lexical class, verbs, function words and adjectives being learned later than nouns. PC2 interacts with verbs, which are learned later than nouns. PC3 and PC6 each interacted with function words, which are learned later than nouns.

For French, PC1 and PC2 predict the age of acquisition. Words which are frequent in general are learned earlier than less frequent words, and both books and speech are important. PC4 interacts with adjectives, PC5 with function words and PC6 with other. In both languages, there is no significant effect of word length.

To sum up, in both languages, we observe that the predictive value of word frequency comes from both speech and books. However, for English, the difference across registers (difficulty of adult-directed versus child-directed) is emphasized. For French, the difference and difficulty of text, independent of register, is emphasized.

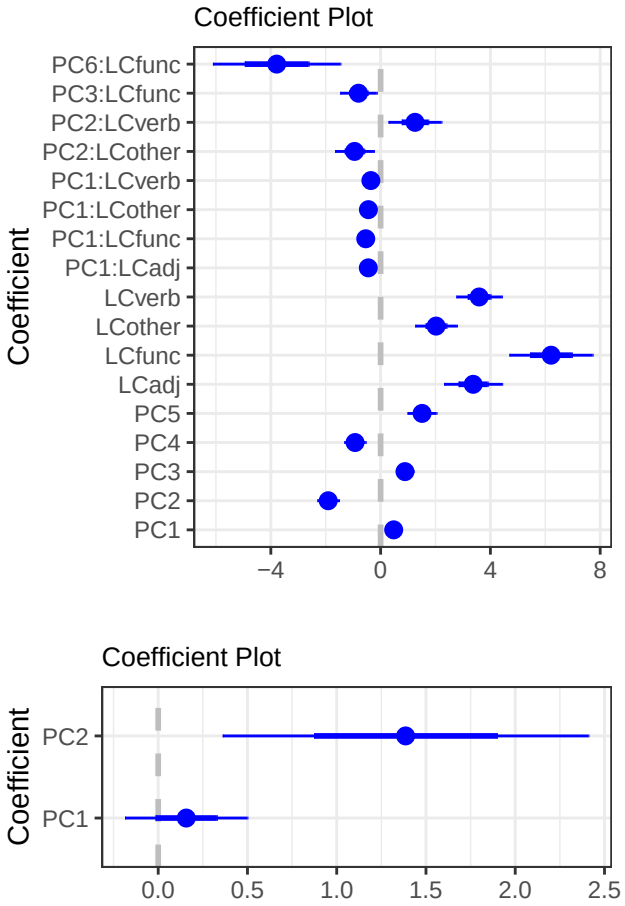

Figure 2: Regression coefficients for predicting CDI AoAs with PCs and lexical class for English (top) and French (bottom).

Principal components and maternal education (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). These findings suggest 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 sec-

[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.

ondary 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 3.

	Beta	SE	t-val	p-val
(Intercept)	-5.33	0.05	-117.54	0.00
age	0.21	0.00	204.85	0.00
MotherEdCollege	-1.54	0.03	-53.74	0.00
MotherEdGraduate	-1.76	0.03	-55.35	0.00
PC1	0.07	0.01	7.39	0.00
PC2	0.31	0.04	7.54	0.00
PC3	-0.29	0.05	-6.07	0.00
PC4	0.15	0.05	2.81	0.01
age:MotherEdCollege	0.07	0.00	59.18	0.00
age:MotherEdGraduate	0.09	0.00	67.84	0.00
age:PC1	-0.00	0.00	-7.94	0.00
age:PC2	-0.01	0.00	-6.89	0.00
age:PC3	0.00	0.00	3.92	0.00
age:PC4	0.00	0.00	3.03	0.00

Table 4: Regression coefficients for predicting English CDI items using PCs and mother's education.

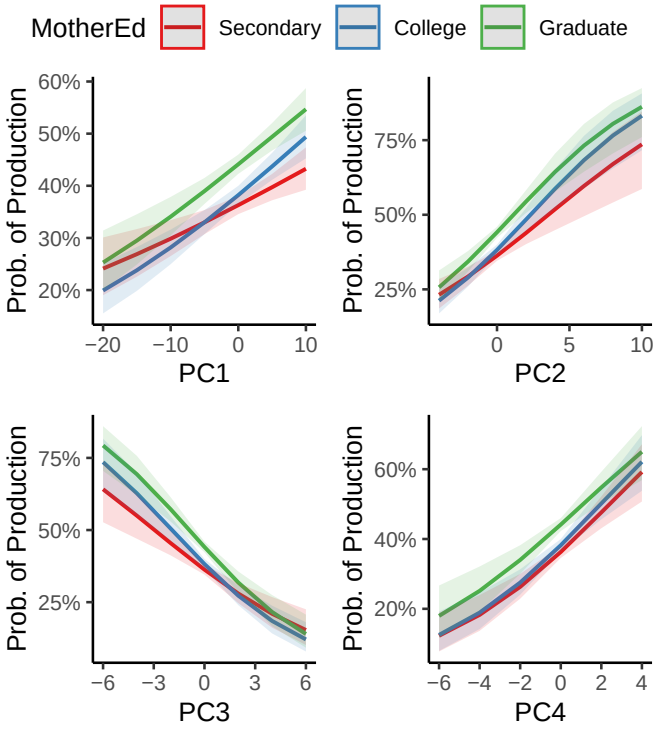

Figure 3: Predicted effects of maternal education and PC1-PC4 on the probability of children producing CDI words.

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 PC1 explains 95% of the variance in frequency distributions. This means that most frequency is shared by the different sources, thus accounted for by that one component. Moreover, multiple components are predictive of children's age of acquisition of words from the CDI. Speech, but also books and media are all relevant in predicting age of acquisition, as well as child-directed speech. This suggests that children's environmental exposure to different language sources, even when these include books or even media, could impact their word learning trajectory. It is somewhat surprising that even sources of input that young children rarely encounter (e.g., adult-directed books) contribute significantly to predicting variation in children's early word learning.

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 con-

texts, 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.