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 in 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. Larry 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, Hsiao, Wei Ming Tan, Banerji, & Nation (2021)].

In this paper, we have three primary research questions. Question 1: 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.

Question 2: 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 (Larry 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, Dale, & Li, 2008).

Question 3: 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.

Together, the answers to these questions provide insight into whether word frequency acts as a single factor in vocabulary learning, or whether different sources and registers have distinguishable effects.

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 approximately 2,400 dyadic telephone conversations. After cleaning, the English AdS yielded 3104000 tokens across 27479 word types. 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). The French AdS yielded 1466000 tokens across 14486 word types.

Adult-directed Books (AdB). The English AdB corpus is taken from a sample of 1,000 Project Gutenberg books tokens randomly selected by Charlesworth, Yang, Mann, Kurdi, & Banaji (2021), totaling 40252700 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 compiled by Charlesworth, Yang, Mann, Kurdi, & Banaji (2021) from online transcripts of movies and TV shows dating from the 1960s (e.g., *Doctor Who*) through the present (e.g., *Breaking Bad*). 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 was drawn from the CDIs (L. Fenson et al., 2007), aggregated in the Wordbank database (Frank, Braginsky, Yurovsky, & Marchman, 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. We computed the proportion of children at each age who were reported to produce each word on the CDI forms completed by parents. We then fit a curve to these proportions using a logistic regression model and determined when the predicted acquisition curve crossed 0.5 (when at least 50% of children produce the word) .

Merging the Corpora

We focus our analysis on the 670 words from the English CDI and 632 words from the French CDI that were present in at least one of the corpora. For French, because of the presence of more complex morphology, CDI words were matched to related words in corpora via a stemmer. All word frequencies were normalized to number of tokens per million (TPM). 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). Unsurprisingly, there were strong correlations

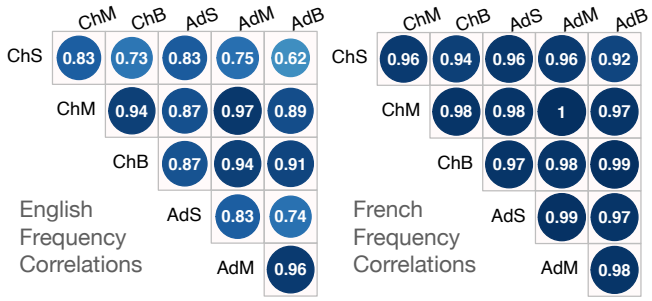

Figure 1: Word frequency correlations between different corpus sources for the matched CDI words in English (left) and French (right).

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, Yurovsky, Marchman, & Frank, 2019; Goodman, Dale, & Li, 2008), we perform our PCA over log frequencies.

Table 1 provides qualitative descriptions of the principal components (PC1-PC6) for each language. The eigenvectors of the PCs in relation to the original six frequency distributions are summarized for English in Table 2, and for French in Table 3, along with the proportion of variance (PVar) explained by each component. 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. Note that the signs of eigenvectors in PCA are arbitrary, and so for ease of interpretation, whever possible we describe findings in the positive direction.

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.

Table 1: Descriptions of English and French PCs based on their order of importance, as well as the words with the highest and lowest values within each PC. CDS stands for child-directed speech. ADS stands for adult-directed speech. CDB stands for child-directed books. CDM stands for child-directed media. ADM stands for adult-directed speech. ADB stands for adult-directed books. EN stands for English and FR stands for French. 'le' = 'the', 'ça' = 'this', 'lequel' = 'which', 'salut' = 'hi', 'madeleine' = 'small cake', 'grand-mère' = 'grandmother', 'doigt de pied' = 'toe', 'sombre' = 'dark', 'parce que' = 'because', 'carottes' = 'carrots'.

Lang	PC	Description	Highest	Lowest
EN	1	overall freq	play dough	the
EN	2	CDS	don't	tissue
EN	3	ADS	don't	gotta
EN	4	Media/Book	camera	was
EN	5	CDS/CDB	grrr	mommy
EN	6	CDM/ADM	beans	don't
FR	1	overall freq	doigt de pied	le
FR	2	Book/Speech	sombre	ça
FR	3	CDS/ADS	élephant	lequel
FR	4	Media/Speech	parce que	salut
FR	5	CDM/ADM	grand-mère	madeleine
FR	6	CDB/ADB	carottes	grand-mère

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, Yurovsky, Marchman, & Frank, 2019; Goodman, Dale, & Li, 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 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

Table 2: English PC rotations and the proportion of variance (PVar) for each PC. The lowest values have black cells, and the highest values have orange cells.

	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

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 included 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 producing each word [within each language, this predictor is extremely correlated with the number of phonemes and serves as a good proxy for production complexity; (**braginsky2018consistency?**)]. 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 ($R^2 = .58$). For French, the model which only included PC1 and PC2 as predictors was significantly preferred, even though the french model explained less variation of the dependent value overall ($R^2 = .06$). 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.

For English, PC1, PC2, PC3, PC4 and PC5 significantly predict the age of acquisition. Overall frequency (PC1) is a predictor; words which are frequent in general are learned earlier than less frequent words (recall that eigenvectors on PC1 are negative and so a positive coefficient indicates greater frequency predicts earlier learning). Child-directed speech (PC2) is a predictor; words which are frequent in this register are learned substantially earlier (more so than for general frequency). On the contrary, for the adult-directed speech predictor (PC3), frequent words are learned later. Word frequency in media distinguished from words in books is a predictor (PC4); words which are frequent in media tend to be learned earlier on. Word frequency in child-directed speech as distinguished from words in child-directed books is also a predictor (PC5), 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, PC2 significantly predicts the age of acquisition, implying the importance of both speech and book sources in explaining variation. In general, the PC1 coefficient direction indicates that frequent words are learned earlier than less frequent words, but it is not a significant predictor in this regression. This could be attributed to PC1 being explained away by PC2, or it is an artifact of the data e.g. some lexical category being more represented than others). PC1 interacts significantly with function words. Verbs are also a significant predictor. In both languages, there is no significant effect of word length.

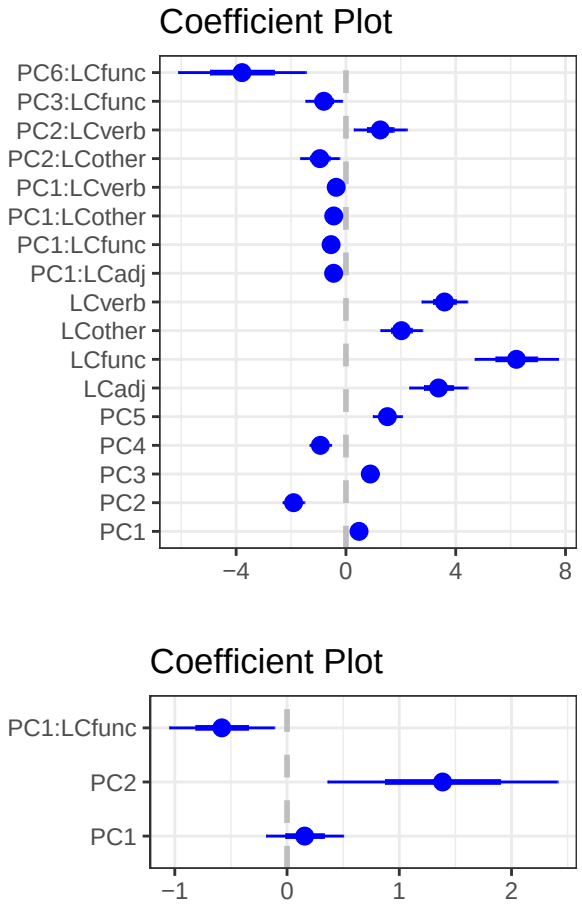

Figure 2: Significant regression coefficients for predicting CDI AoAs with PCs and lexical class for English (top) and French (bottom). 'func' stands for function words, 'other' stands for words refering to games, routines and people. 'adj' stands for 'adjective'.

To sum up, in both languages, we observe that the predictive value of word frequency was especially high for words that are highly frequent in child directed speech However, for English, the difference across registers (difficulty of adult-directed versus child-directed) appears more important. For French, the difference and difficulty of text, independent of register, appears more important.

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 whose mothers are more highly-educated (a proxy for household SES) 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 fit a series of exploratory logistic regressions successively adding the PCs to predict the number of children in Wordbank who produce or don't produce each item. Due to lack of maternal education data for French data, we focus on American English-learning children. We included interactions of mother's education and children's age with each included PC; interactions of this type indicate that a particular type of register or source might be more important to acquisition for one SES group or another.

American English data from Wordbank contained 2,776 CDI:WS administrations with mother's education dichotomously coded (N=1160 with at most some college education; N=1616 with a college degree or more). The series of ANOVAs indicated that PC1 through PC5 significantly improved the model fits, but that adding PC6 was not justified. Thus, we analyzed the model that included the first five PCs.

This model showed significant main effects of age, mother's education, and all five PCs (all $p < .001$). Faster learning was predicted for words with higher values on PC1 (overall frequency; $\beta = .10$), PC2 (child-directed speech; $\beta = .45$), or PC4 ($\beta = .28$). Slower learning was predicted for words with higher values on PC3 ($\beta = -.34$) or PC5 ($\beta = -.49$).

There was a significant positive interaction of age with mother's education ($\beta = .06$, $p < .001$), shown in Figure 3. There were also significant interactions of age and all five PCs, although the coefficients were all of a small magnitude (βs $< .01$). Only PC1 and PC5 interacted significantly with mother's education. Shown in Figure 3, children of higher-educated mothers were more likely to know words that were higher on PC1 (overall frequency; $\beta = .03$, $p < .001$), while they were less likely to know words that were high on PC5 ($\beta = -.09$, $p = .01$). Words more frequent in child-directed speech are high on PC5, while words that are more often in child-directed books are low on PC5, meaning that this interaction indicates children with higher-educated mothers are more likely to learn booky (rather than speechy) words.

There was one significant 3-way interaction of age, mother's education, and PC1, with a small magnitude negative coefficient ($\beta = -.001$).

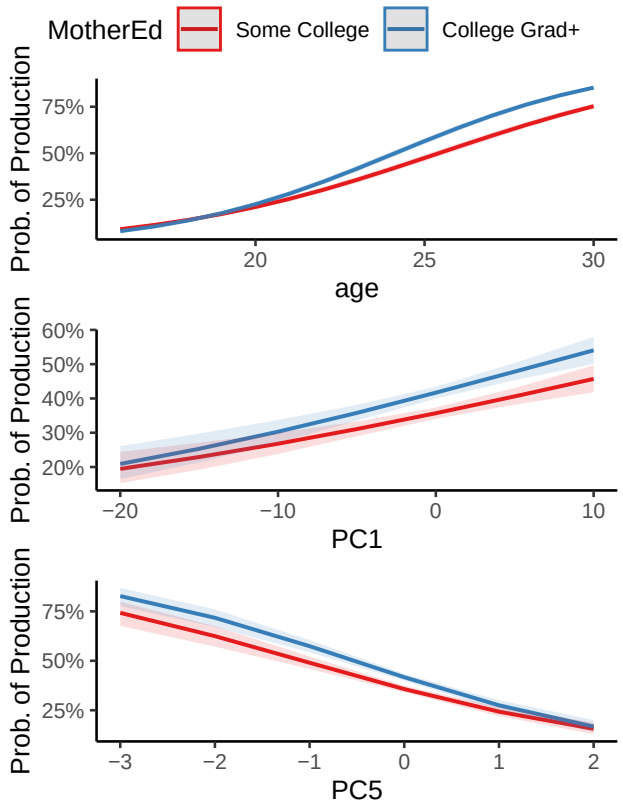

Figure 3: Predicted effects on the probability of English-speaking children producing CDI:WS words based level of maternal education. A significant positive interaction with age (top) shows an increasing effect of maternal education as children age. A significant positive interaction with PC1 (middle) shows that words with higher overall frequency are produced more by children with more educated mothers. Finally, a negative interaction with PC5 shows that children with more educated mothers are likely to produce words more representative of child-directed books, rather than child-directed speech.

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 90% of the variance in frequency distributions for both languages. This means that most frequency is shared by all the different sources and registers in our study, and is thus accounted for by this one component. In spite of this, the remaining prinicipal components in both languages picked up on consistent differences in both register and source. In English, child-directed speech captures a large part of the variance as a second principal component, which shows that CDI word frequencies actually differ a lot compared to all other sources. In French, in contrast, child-directed speech loaded strongly on the third principal component, while the second component picked up on bookish words (both child- and adult-directed) – a distinction captured in PC4 for English.

Multiple components are predictive of children's age of acquisition of words from the CDI in English. Child-directed speech, adult-directed speech, but also books and media are all relevant in predicting age of acquisition. 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, and 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.

In French, the component diversifying books and speech captures a large part of the variance as a second principal component, and is the most relevant one in predicting age of acquisition. This finding underlines the importance of the speech and text sources independent of register. On the contrary, the difference between child-directed and adult-directed registers is not as pronounced as in our English data. We speculate that the French word frequencies tested here do not differ that much across registers and instead vary more between spoken and written sources.

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 the first five PCs, as well as interactions of mother's education with PC1 (overall frequency) and PC5 (child-directed speech vs. books) – but without any significant interaction of PC2 (child-directed speech), nor PC3 (adult-directed speech), nor of the media component (PC4). This finding suggests that the early language advantage shown by children of more highly-educated mothers (and thus in higher-SES households; cf. Hoff, 2003) may in part be due to greater amounts of shared reading time.

This last finding is especially interesting because it suggests the specific role of literacy practices in affecting vocabulary growth. 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]).

This research has a number of limitations that point the way to future work. First, our study is an observational linkage between frequencies as estimated from one set of materials and acquisition trajectories from wholly different children. We expect that frequencies represent estimates of an

average experience by a member of a particular linguistic or cultural group, but they are certainly biased by their specific source. Further, 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 makeup, e.g. small stories in French vs. large books in English. Finally, although we made an effort to examine two languages, French and English represent only a tiny subset of the broader set of linguistic environments in which children acquire their vocabulary.

In sum, 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

10 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, Larry, 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.