Identifying the distributional sources of children's early vocabulary

George Kachergis*, Georgia Loukatou*, and Michael C. Frank

kachergis, loukatou, mcfrank@stanford.edu

Department of Psychology, Stanford University, Stanford, CA 94305 USA
*equal contributions

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 input sources – in both English and French – and to evaluate how predictive these distributions 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 household socio-economic status (SES) 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- vs. adult-directed – of the language. These input sources vary in word frequency, as well as by various measures of quality. For example, children's books have higher lexical diversity than child-directed speech, and thus may represent an important source of lexical knowledge (Montag, Jones, & Smith, 2015). 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 (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 household SES (Hoff, 2003; Rowe, 2012). SES is a composite concept and parental education has often been used as a proxy for SES (see Rowe, 2018 as an entry point to this literature). For example, Hoff (2003) compared the speech of low- versus high-SES American mothers, with SES defined based on education (college-educated versus high school).

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 individual words' age of acquisition (Goodman, Dale, & Li, 2008). But 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 et al., 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 primary research questions. Question 1: How different are different input sources? 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) as a way to understand the relation between frequency distributions from different sources and registers.

Question 2: What is the relation between input frequencies and acquisition? 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 include parent-report checklists measuring children's early vocabu-

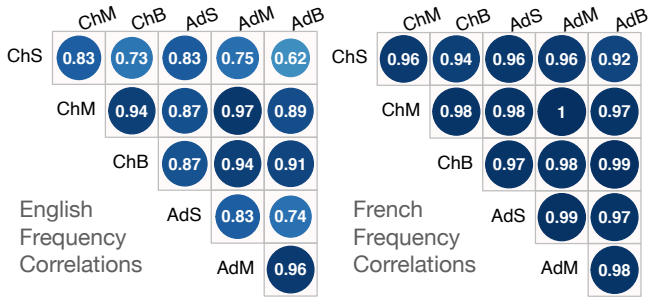

Figure 1: Word frequency correlations between different corpus sources for the CDI words in English and French.

lary, which have proven to be reliable and valid indicators of children's growing language skill (Fenson et al., 1994). Critically, CDI forms provide information about >600 individual words that children eventually learn to 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 the mean age (in months) at which 50% of children are expected to know a given CDI word (Braginsky, Yurovsky, Marchman, & Frank, 2019; Goodman et al., 2008).

Question 3: How do input frequencies relate to maternal education? An intuitive hypothesis is that children of more highly-educated parents may read more to their young children, although it should be noted that parental education is strongly associated with household SES–to the extent that the former is often used as a proxy for the latter. Indeed, young children from high-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. Thus, we test which sources of input frequency combine with maternal education to better predict children's acquisition of particular words, expecting that we may see evidence of earlier learning of words from children's books in households with higher maternal education (and 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

We used corpora from different sources 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, together with all analysis scripts.

Child-directed Speech (ChS). Utterances of ChS were extracted from CHILDES (MacWhinney (2000); excluding book reading corpora), a collection of transcripts of interactions between caregivers and children 0 to 12 years of age ($M = 2.9$ years). After cleaning, the CHILDES English corpus yielded 5521000 tokens across 38779 word types. The French ChS yielded 3190000 tokens across 13139 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 6724000 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 et al. (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 et al. (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 767000 tokens across 15635 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 (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. 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 are reported to 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 (Porter, 2001). 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.[1]

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. Overall, French distributions were more highly correlated with one another, likely due to smaller corpus sizes.[2]

We used principal components analysis 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 PCAs over log frequencies. The eigenvectors of the PCs in relation to the original six frequency distributions are summarized for English in Table 1, and for French in Table 2, along with the proportion of variance (PVar) explained by each component.

[1]Other forms of smoothing, e.g. Laplace smoothing ($\alpha = 10$; added to all counts, including missing words) yielded similar results.

[2]The French corpora had a large number of CDI words missing (and thus smoothed): 95 in AdB; 77 in AdS; 70 in AdM; 99 in ChB; 12 in ChS; 53 in ChM. In comparison, English had 1 missing in ChB, ChS, and ChM; 14 in Adb; 45 in AdS; 28 in AdM.

Table 1: English PC rotations and proportion of variance (PVar), colored by value (low=black; high=orange).

	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 2: French principal component rotations.

	PC1	PC2	PC3	PC4	PC5	PC6
ChM	*-0.41*	-0.09	0.06	**-0.65**	**0.54**	0.32
ChB	*-0.41*	**0.55**	0.20	0.13	0.28	**-0.63**
ChS	*-0.36*	**-0.50**	**0.73**	0.24	-0.16	0.01
AdM	*-0.42*	-0.19	-0.33	-0.42	**-0.61**	-0.35
AdB	*-0.41*	**0.54**	0.01	0.18	-0.36	**0.62**
AdS	*-0.42*	-0.34	**-0.56**	**0.54**	0.30	0.05
PVar	0.90	0.04	0.03	0.02	0.01	0.00

PC1 already explains the bulk of the variance (89% for English and 90.2% for French); and PC2-PC4 each only capture an additional 2-4% of the variance for both languages.

Table 3 provides qualitative descriptions of the principal components (PC1-PC6) for each language. Note that the signs of eigenvectors in PCA are arbitrary, so on some PCs positive values correspond to higher absolute frequencies, while on others negative values correspond to higher absolute frequencies. 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. For English, 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; it mostly captures child-directed speech for English, whereas for French it captures book language.

PCA-based Age of Acquisition Regression (Q2)

Next, we turned to the question of how well these frequency distributions predict English- and French-learning children's early word learning. ChS has been claimed to present properties that could facilitate language acquisition and promote infants' attention to language (Golinkoff, Can, Soderstrom, & Hirsh-Pasek, 2015; Soderstrom, 2007) when compared to

Table 3: English (EN) & French (FR) PC descriptions.

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

Note. PCs ordered by importance, and example words with highest and lowest values on each PC. CDS = child-directed speech; ADS = adult-directed speech; CDB = child-directed books; CDM = child-directed media; ADM = adult-directed media; ADB = adult-directed books. FR to EN translations: 'le'='the', 'ça'='this', 'lequel'='which', 'salut'='hi', 'madeleine'='cookie', 'grand-mère'='grandmother', 'doigt de pied'='toe', 'sombre'='dark', 'parce que'='because', 'carottes'='carrots'.

AdS. Frequency distributions in ChS could thus play a role in predicting children's early word learning. There has been less evidence on the specific role of books and media in predicting 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. To test this, we ran a regression predicting AoA with log(word frequency) from each of the six distributions as predictors, and examined the Variance Inflation Factor (VIF), which measures how much a regression coefficient variance is inflated 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 to predict AoA instead of the raw frequency distributions.

As past research indicates that lexical class strongly modulates influences of word frequency, we included the two-way interaction of lexical class (LC) with each PC 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 highly correlated with the number of phonemes and serves as a good proxy for production complexity; Braginsky et al. (2019)]. 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[3]. 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 variable overall ($R^2 = .06$). Figure 2 shows the significant coefficient estimates ($p < 0.05$) for English, and all main effects for French, as well as the one significant interaction.

For English, PC1, PC2, PC3, PC4 and PC5 significantly predict the age of acquisition. Overall frequency (PC1) is a predictor; frequent words 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; frequent words 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 finding could be attributed to PC1 being explained away by PC2, or it could be an artifact of the data e.g. some lexical category being more represented than others. PC1 also interacted significantly with function words. In both languages, there was no significant effect of word length, unlike in prior studies; this finding may indicate that prior effects of word length were confounded with register or source frequency effects (Braginsky et al., 2019).

Table 4 shows the top 5 words with the most-improved AoA prediction (greatest decrease in residual squared error) with the addition of each particular PC as a regression predictor for both languages. These show qualitative correspondence with the interpretations of the PCs shown in Table 3. For example, when PC2 – roughly corresponding to child-

[3]R syntax for the sequence of regressions (noun as baseline lexical category): `AoA~PC1*LC`, `AoA~(PC1+PC2)*LC`, ..., `AoA~(PC1+PC2+PC3+PC4+PC5+PC6)*LC`

directed speech – is added to the English model, "uh oh" and "no" improve; when PC5 – roughly corresponding to child-directed books – is added, several animal noises like "grrr" and "cockadoodledoo" improve.

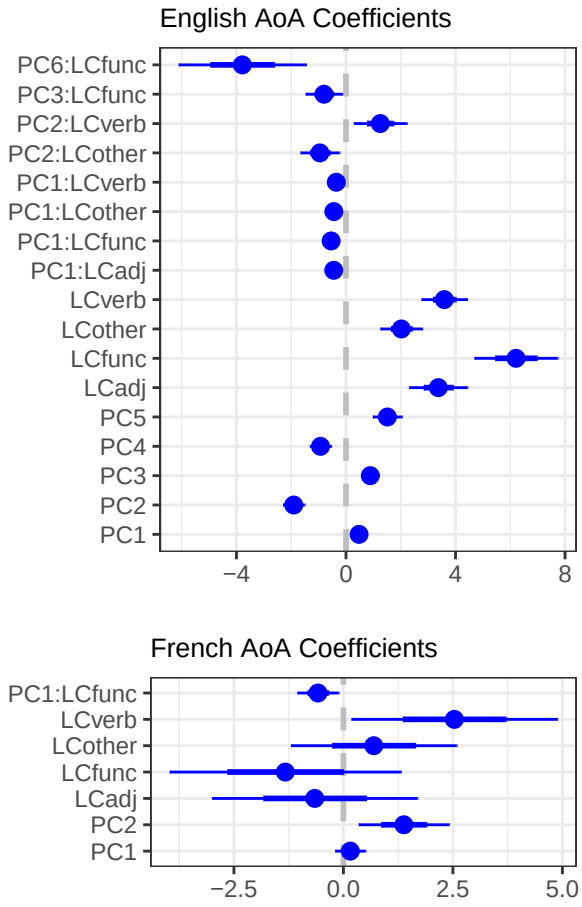

Figure 2: Significant regression coefficients for predicting CDI AoAs with PCs and lexical class for English (top) and French (bottom). Abbreviated levels of lexical class (LC) are function words ('func'), adjectives ('adj'), and 'other' includes items refering to games, routines, and people.

Principal components and maternal education (Q3)

Correlations between household SES and children's language development are documented in a large body of literature, and SES seems to be predictive of aspects of word learning (Hoff, 2003). Previous findings also relate SES status to book reading (Shen & Del Tufo, 2022), which in turn seems to be related to better language skills (Bus, Van Ijzendoorn, & Pellegrini, 1995; Sénéchal & LeFevre, 2002). These findings suggest that the vocabulary composition of children whose mothers are highly-educated (parental education being 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

Table 4: Top 5 words with improved prediction of AoA when adding each PC for English (EN) and French (FR)

Lang PC	Top 5 Words
EN +1	can, no, cockadoodledoo, now, time
EN +2	yes, gas, don't, uhoh, no
EN +3	camping, bye, jeans, smile, babysitter
EN +4	was, lot, lips, gonna, hafta
EN +5	mommy, grrr, cockadoodledoo, tissue, babysitter
EN +6	would, mine, does, my, could
FR +1	le, faire, au sommet de, au sujet de, lequel (EN: the, do, on top, about, which)
FR +2	coincé, sombre, maître, oui, aïe (EN: stuck, dark, teacher, yes, ouch)

or don't produce each item. Due to lack of maternal education data for French data, we focused on American English-learning children. We included interactions of mother's education and children's age with each PC; such interactions 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; $\beta = .45$), and PC4 (media vs. books; $\beta = .28$). 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 $< .01$).

In the critical test of our hypothesis, 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 words that occur in books (rather than words that occur in speech). As seen in Table 4, these include animal sounds, which occur frequently in baby books and in other analyses tend to be known more by children whose mothers have more education (Frank, Braginsky, Yurovsky, & Marchman, 2021).

General Discussion

We investigated how linguistic input varies across the types of language that children may experience using word frequency distributions garnered from child-directed and adult-directed corpora of speech, books, and media (TV and movies). Since frequencies are highly correlated, we found the principal components (PCs) of these distributions, which revealed systematic variation along source and register dimensions. Most variation in frequency is shared across all the different sources and registers in our study. French in particular showed high correlations, perhaps due to smaller corpus size or due to noise introduced by stemming and lemmatization. In spite of this, other PCs in both languages picked up on consistent differences in both register and source. In English, child-directed speech captured a substantial part of the variance as a second principal component, suggesting real differences in word frequencies in this register. 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 by a smaller PC for English.

Multiple components were predictive of children's age of acquisition of words from the CDI, especially for English. Child-directed speech, adult-directed speech, but also books and media were all relevant in predicting age of acquisition. In French, the component distinguishing books and speech was most relevant in predicting age of acquisition. Lending some qualitative support to these conclusions, the specific words that were improved by the addition of particular PCs appeared somewhat related to these sources and registers.

We also used English Wordbank data to test how well frequency components combine with mother's education to predict early word learning. This exploratory 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). This intriguing finding suggests that the early language advantage shown by children of highly-educated mothers (and thus in higher-SES households; cf. Hoff, 2003) may in part be due to greater amounts of shared reading time. Future research may predict individual children's learning of particular words using these principal components, in combination with measures of the time their child spends receiving input from each input source ([adult- vs. child-directed] x [books, media, speech]).

This research has a number of limitations that point the way to future work. First, it 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. Second, corpora of different sizes represent the sources and registers for each language, and less data were available overall for French. Observed differences between the two languages could thus be partly attributed to differences in corpus size and sources. For example, the French corpora are composed of several smaller ones and have slightly different makeup, e.g. French short stories vs. English longer books. Third, most child books were published decades ago, and are used due to the lack of open-source contemporary books. This may yield different frequency distributions, but the difference should be negligible for the CDI words tested here. Moreover, we note the difficulty of drawing conclusions about these results, since they are based on interpretations of the relevant dimensions of the different principal components. Finally, although an effort was made to examine two languages, they 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 that 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.

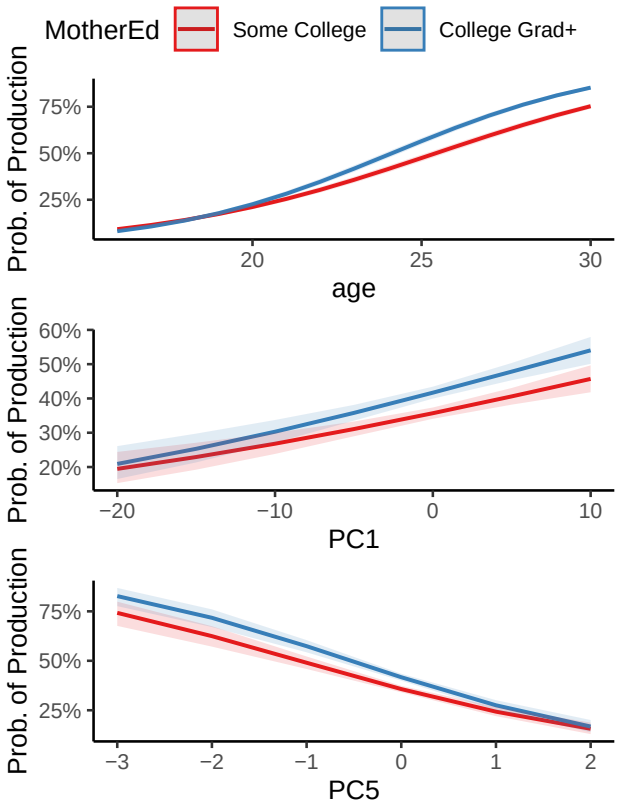

Figure 3: Predicted effects on the probability of English-speaking children producing CDI:WS words based on maternal education. A significant positive interaction with age (top) shows an increasing effect of maternal education as children age. A significant positive interaction with PC1 (middle) shows that words with higher overall frequency are produced more by children with more educated mothers. 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.

Acknowledgements

This research was supported by the Stanford Maternal and Child Health Research Institute, and funded in part by the Fyssen Foundation. We thank members of the Language and Cognition lab for their feedback.

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, 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.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The wordbank project*. MIT Press.

Godfrey, J., & Holliman, E. (1993). Switchboard-1 release 2 LDC97S62. *Linguistic Data Consortium*.

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby) talk to me: The social context of infant-directed speech and its effects on early language acquisition. *Current Directions in Psychological Science*, *24*(5), 339–344.

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.

Porter, M. F. (2001). Snowball: A language for stemming algorithms.

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.

Rowe, M. L. (2018). Understanding socioeconomic differences in parents' speech to children. *Child Development Perspectives*, *12*(2), 122–127.

Sénéchal, M., & LeFevre, J.-A. (2002). Parental involvement in the development of children's reading skill: A five-year longitudinal study. *Child Development*, *73*(2), 445–460.

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.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532.