Identifying the distributional sources of children's early vocabulary

Anonymous CogSci submission

Abstract

Children's early vocabulary learning must to a large extent be driven by the prevalence of words: they can't learn a word if they haven't heard it. Indeed, previous research has found that higher word frequency is a good predictor of earlier learning. However, despite considerable overlap, word frequency distributions also vary significantly by source: child-directed speech, books, and television have distinct profiles. Children receive a mixture of these different frequency distributions, and the ratios of the mixture may be predictive of their early word learning. The goal of this paper is to better understand the shared and unique variance in these sources of input–in both English and French–and to evaluate how predictive these input frequencies are of children's early word learning.

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

Introduction

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). However, input word frequency varies significantly depending on the context. For instance, word frequency in books is not the same as frequency in conversational 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 (2,260 tokens per million; TPM), yet not so common in children's books (10 TPM), and even more rare in books meant for all ages (2 TPM). But other differences are less intuitive: "of" is frequent in books meant for all ages (41,630 TPM), and while still frequent in child-directed speech (5,900 TPM), relatively less so as compared to children's books (20,400 TPM). In general, speech–both directed to children, and to adults–contains relatively fewer function words, and tends to score lower on measures of lexical diversity than books (Dawson et al., 2021).

Different language input sources, such as the use of child-directed register or book reading time, can lead to variance in input speech heard by child during their everyday lives. This input variance has often been interpreted as a function of the families' socioeconomic status (SES; Rowe, 2018). Importantly, this variance has been found to relate to children's language development and to be predictive of aspects of word learning (Hoff, 2003).

The primary goal of this paper is to examine shared and unique variance in word frequency across different sources of input, ranging from children's books and movies to child-directed speech and adult-directed books, movies, and speech, which we accomplish via principle components analysis (PCA). After characterizing the structure of the principle components of frequency from different sources in both English and French, we investigate how well these components predict children's early word learning, using aggregate MacArthur-Bates Communicative Development Inventories (CDI) data from Wordbank (Frank, Braginsky, Yurovsky, & Marchman, 2017). Finally, we examine how well the frequency components predict individual children's word learning in combination with their mother's education, which may be related to how much children are read to at home.

Method

Datasets

Child-directed Speech. Utterances of child-directed speech (ChS) were extracted from the CHILDES corpus (MacWhinney, 2000), a collection of transcripts of dyadic interactions between caregivers and children of ages ranging from 0 to 12 years ($M = 2.9$ years). After cleaning, the CHILDES corpus yielded a total of 5.521096×10^6 tokens across 38779 word types.

Child-directed books (ChdB). We used a sample of 98 children's books from Project Gutenberg's open-source database of books that has been used in prior machine learning research on language comprehension (Hill, Bordes, Chopra, & Weston, 2015). These books were published between 1820 and 1922, but include such well-known titles as *The Legend of Sleep Hollow*. After cleaning, this children's book corpus totals 4.674202×10^6 tokens across 42666 word types.

Child-directed Media (ChM). Transcripts from 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 were taken from Charlesworth, Yang, Mann, Kurdi, & Banaji (2021) (available here: `https://osf.io/kqux5/`. After cleaning, this children's media corpus totals 6.759247×10^6 tokens across 84333 word types.

Adult-directed Speech (AdS). Adult-directed speech was obtained from the Switchboard-1 Telephone Speech Corpus (Godfrey & Holliman, 1993), a corpus of transcripts from dyadic telephone conversations in which 543 adult speakers were assigned to discuss a randomly-assigned topic. After cleaning, the adult-directed speech corpus (AdS) yielded a total of 3.104328×10^6 tokens across 27536 word types.

Adult-directed Books (AdB). The adult-directed book corpus (AdB) is comprised of 3.5143617×10^7 tokens across 827414 word types.

Adult-directed Media (AdM). The adult-directed media corpus (AdM) is comprised of 6.166909×10^6 tokens across 62876 word types.

Merging the Corpora

Children's early word learning data is drrawn from the CDIs (Fenson et al., 2007). CDIs are parental reports on their children's lexical development, proven to be reliable indicators of a child's language. CDIs are survey instruments, where parents mark whether their child (age ranges 8-15 and 16-30 months old) understands or produces particular words out of a list of several hundred words.

All word frequencies were normalized to number of tokens per million (TPM). We focus our analysis on the 674 words from the CDI that we were able to find in at least some of the corpora. We were unable to match 6 CDI items in any of the corpora, including "babysitter's name", "child's own name", ... 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 missing word was replaced in each of ChB, ChM, and ChS, while 14 missing words were replaced in AdB. ...

Results

Cross-corpus Frequency Correlations

Table 1 shows the cross-corpus word frequency correlations for the 674 CDI words.

Table 1: Correlation table of word frequency distributions for CDI words.

	ChS	ChM	ChB	AdS	AdM	AdB
ChS	1.00	0.83	0.73	0.83	0.75	0.62
ChM	0.83	1.00	0.94	0.87	0.97	0.89
ChB	0.73	0.94	1.00	0.87	0.94	0.91
AdS	0.83	0.87	0.87	1.00	0.83	0.74
AdM	0.75	0.97	0.94	0.83	1.00	0.96
AdB	0.62	0.89	0.91	0.74	0.96	1.00

Age of Acquisition Regression

Despite the strong correlations of word frequency across these different corpora, we will attempt a simple regression predicting each CDI word'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). We also include the number of letters as a predictor (Nletters) to help control for the overall difficulty of each word. Table 2 below displays the estimated coefficients of this regression, showing significant contributions of all child-directed distributions (speech, books, and media), and of adult speech frequencies, as well as the number of letters. However, the variance inflation factor (VIF) values for all of the frequency distributions are all $>> 1$ (and many > 5 or > 10), indicating that these variables show strong multicollinearity which may compromise the reliability of the regression results. Thus, we turned to principal components analysis (PCA) to disentangle these correlated distributions and to understand their interrelations.

	Beta	SE	t-val	p-val	VIF
(Intercept)	27.39	0.90	30.44	0.00	
log(ChS)	-2.16	0.17	-12.37	0.00	4.56
log(ChM)	-0.52	0.23	-2.24	0.03	9.46
log(ChB)	1.42	0.17	8.18	0.00	6.64
log(AdS)	0.59	0.12	5.07	0.00	6.34
log(AdM)	0.18	0.18	1.00	0.32	12.63
log(AdB)	0.18	0.17	1.04	0.30	11.96
Nletters	0.40	0.09	4.71	0.00	1.45

Table 2: Coefficients for corpora frequencies predicting AoA.

Principal Components of Frequency

We use PCA to examine the principal components of the six log-scaled word frequency distributions ([adult- vs. child-directed] x [speech, books, media]). Table 3 shows the standard deviation (Std Dev) and proportion of variance explained, both individually (Prop Var) and cumulatively (Cum Prop Var) by the principal components (PC1-PC6). PC1 already explains the bulk the variance (89%), and PC2-PC4 each only capture an additional 2-4% of the variance. In total, the first four components capture >98% of the variance.

Table 3: Importance of components from PCA.

	PC1	PC2	PC3	PC4	PC5	PC6
Std Dev	5.10	1.05	0.97	0.82	0.57	0.44
Prop Var	0.89	0.04	0.03	0.02	0.01	0.01
Cum Prop Var	0.89	0.93	0.96	0.98	0.99	1.00

Table 4 shows the eigenvectors of the principal components (PC1-PC6) in relation to the original six frequency distributions. Table 5 shows the correlation of each frequency distribution with the CDI items' loadings on each principal component [GK: cut one of these tables? which?]. It is clear that PC1 captures overall frequency, but loads more strongly on

the adult distributions (-.47 to -.52). PC2 (3.8% of variance) mostly captures child-directed speech (0.65), especially differentiating it from adult-directed books and media (-.47, -.45). PC3 (3.1% of variance) captures adult-directed speech (.79), particularly from child-directed distributions. PC4 (2.2% of variance) captures the similarity of child- and adult-directed media (.40, .48), distinguishing them from child- and adult-directed books (-.63, -.44). PC5 (1.1% of variance) mostly captures child-directed books (0.53), distinguishing it from child-directed speech (-.60) and adult-directed books (-.50). PC6 (<1% of variance) captures child-directed media (.73), distinguishing it in particular from adult-directed media (-.52). In summary, the principal components align surprisingly well with particular dimensions of the frequency distributions: PC1 with overall adult-directed 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. Although even PC5 and PC6 seem interpretable, given limited space and the fact that they account for very little variance (˜1%) we will focus on PC1-PC4 for the rest of our analyses.

Table 4: Principal components' rotation in the original coordinate system.

	PC1	PC2	PC3	PC4	PC5	PC6
ChM	-0.31	0.15	-0.35	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

PCA-based Prediction of Age of Acquisition

Figure 1 shows the loadings of CDI items on PC1-PC4 vs. the average age of acquisition (AoA; in months).

Regularized Regression on PCA Loadings

Next we used the CDI items' PCA loadings in lieu of the frequency distributions to predict AoA. First, we did a L1-regularized (i.e., LASSO) regression predicting AoA with all of the principal components, to test which of the PCs should be included in the regression with lexical class. We used cross-validation to find the λ value (penalty for outsized coefficients) that minimized mean-squared error of the test set, resulting in $\lambda = 0.006$, a small value which will result in coefficients that would be quite close to those obtained in ordinary least squares regression (i.e., when $\lambda = 0$). Table 5 displays the estimated coefficients using this best-fitting λ, showing that all of the PCs have non-zero values. This model has $R^2 = 0.32$.

	s0
(Intercept)	25.11
PC1	-0.17
PC2	-1.13
PC3	0.88
PC4	-1.32
PC5	1.95
PC6	-0.34

Table 5: LASSO AoA regression coefficients.

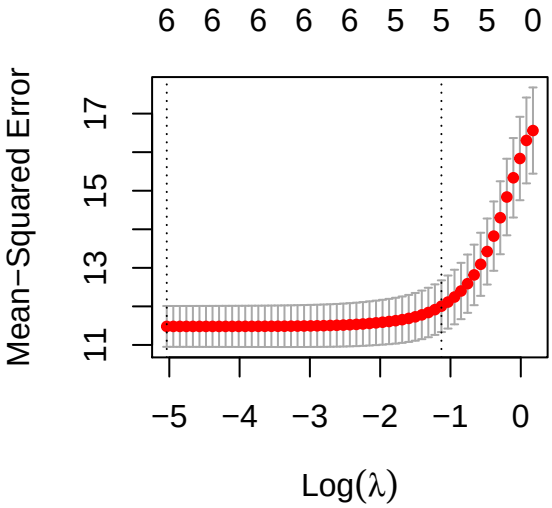

Figure 2: Test MSE for different lambda values in the cross-validated lasso regression. The minimum MSE is achieved by the leftmost dotted line, while the rightmost dotted line shows the lambda that achieves MSE $<$ 1 standard error above this minimum.

Given that past research has found that lexical class strongly modulates influences of word frequency, we next examine the interaction of lexical class (LC) with PC1 - PC6. 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. Thus, the R syntax for the sequence of regressions was `AoA~PC1*LC`, `AoA~(PC1+PC2)*LC`, ..., `AoA~(PC1 + PC2+PC3+PC4+PC5+PC6)*LC`. The more complex model was always significantly preferred, including up to the inclusion of PC6. Table X shows the results of this final regression, which yielded $R^2 = .584$.

Combining distributions with demographic data

Past research has found that young children from higher-SES households tend to have larger vocabulary. Parents with higher-SES tend to also report reading more to their young children than parents with lower SES. Together, this suggests that the vocabulary composition of children from higher-SES households may be better predicted by the word frequencies seen in child-directed books, rather than those from child-directed speech. To test this idea, we did

	Beta	SE	t-val	p-val
(Intercept)	22.38	0.22	99.78	0.00
PC1	0.48	0.05	9.53	0.00
PC2	-1.91	0.20	-9.71	0.00
PC3	0.89	0.16	5.64	0.00
PC4	-0.93	0.19	-4.80	0.00
PC5	1.51	0.26	5.73	0.00
PC6	0.35	0.34	1.02	0.31
LCadj	3.37	0.53	6.37	0.00
LCfunc	6.21	0.76	8.20	0.00
LCother	2.02	0.38	5.33	0.00
LCverb	3.59	0.41	8.66	0.00
PC1:LCadj	-0.45	0.11	-3.91	0.00
PC1:LCfunc	-0.54	0.10	-5.19	0.00
PC1:LCother	-0.44	0.08	-5.38	0.00
PC1:LCverb	-0.35	0.10	-3.44	0.00
PC2:LCadj	0.84	0.49	1.72	0.09
PC2:LCfunc	0.20	0.35	0.56	0.57
PC2:LCother	-0.95	0.35	-2.68	0.01
PC2:LCverb	1.25	0.48	2.60	0.01
PC3:LCadj	-0.52	0.58	-0.90	0.37
PC3:LCfunc	-0.81	0.33	-2.42	0.02
PC3:LCother	0.20	0.35	0.57	0.57
PC3:LCverb	0.31	0.38	0.81	0.42
PC4:LCadj	-0.17	0.66	-0.26	0.80
PC4:LCfunc	-0.17	0.39	-0.43	0.67
PC4:LCother	0.39	0.39	1.00	0.32
PC4:LCverb	0.63	0.50	1.25	0.21
PC5:LCadj	-0.62	0.83	-0.75	0.46
PC5:LCfunc	-0.73	0.62	-1.18	0.24
PC5:LCother	0.55	0.53	1.03	0.30
PC5:LCverb	-0.23	0.74	-0.31	0.76
PC6:LCadj	1.08	1.22	0.88	0.38
PC6:LCfunc	-3.79	1.16	-3.26	0.00
PC6:LCother	-0.48	0.76	-0.63	0.53
PC6:LCverb	-0.92	0.97	-0.94	0.35

Table 6: Regression predicting AoA with PCs and lexical class.

an exploratory MANOVA with 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. The R syntax for the formula is `cbind(total_producing, total_not_producing) ~ age * mother_ed * (PC1 + PC2 + PC3 + PC4)`. American English data from Wordbank contained 2,776 CDI:WS administrations with mother's education (coded: -1 for no more than secondary education (N=547), 0 for some/all college (N=1483), and 1 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 $p < .001$). [how to show these effects?]

Discussion

Acknowledgements

This research was funded in part by (Georgia's grant). We thank members of the Language and Cognition lab for their feedback.

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.

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.

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*.

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.

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.

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

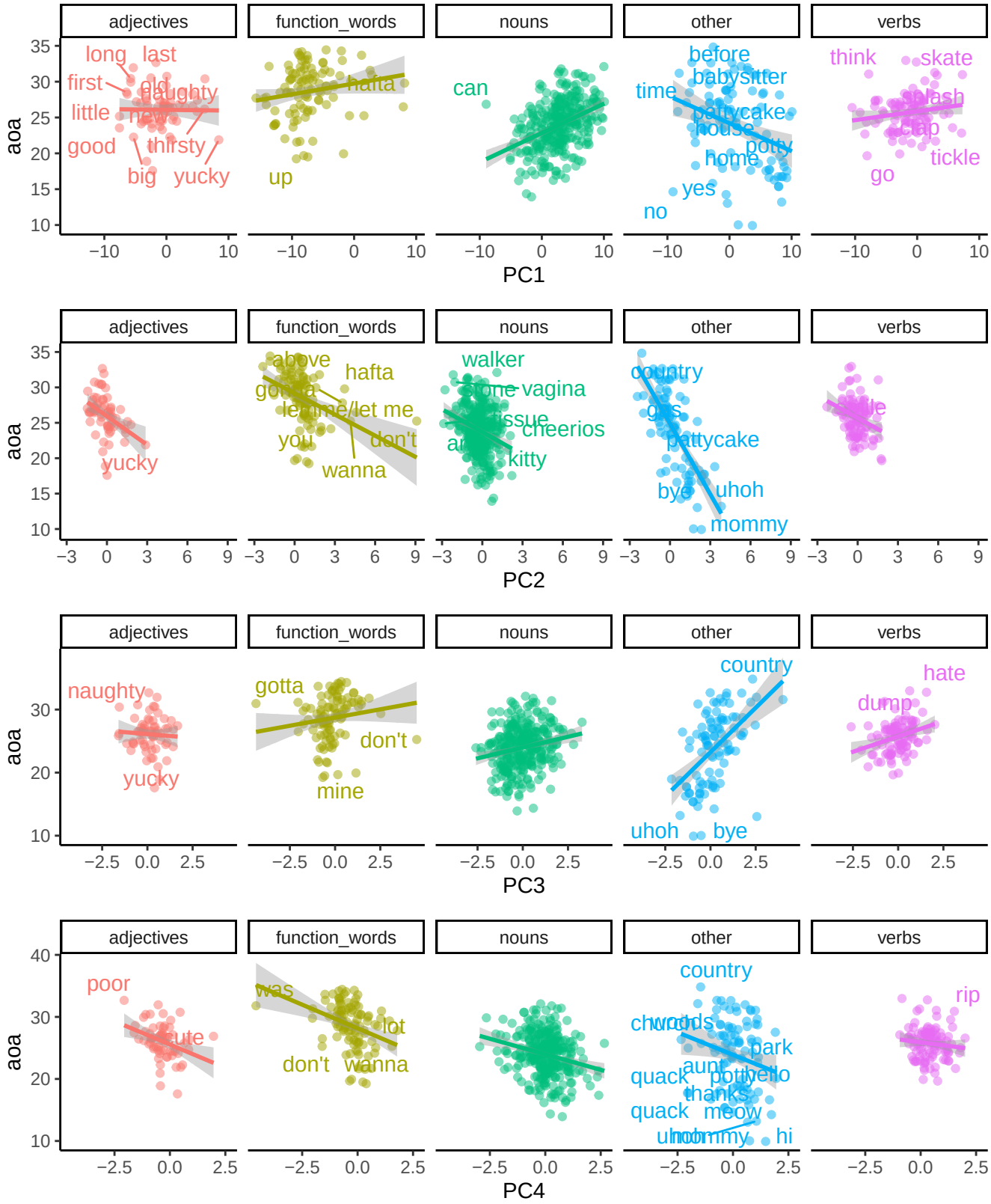

Figure 1: Principal components vs. age of acquisition, by lexical class.