Children's books as a unique source of language input*

Joseph Denby†

June 6, 2018

Abstract

A longstanding research effort within developmental psychology aims to examine early childhood linguistic input and its effects on language development. Some researchers in this vein have employed low-level linguistic analyses to get at mechanistic differences between, e.g., speech from mothers of different socioeconomic statuses. In this paper, I follow a recent trend in the literature by applying computationally-enhanced content analysis techniques to another prominent source of early linguistic input: children's books. I find that the text of children's books is markedly different from child-directed speech: specifically, book text consists of language that is generally more lexically diverse, more complex at a sentence level, and potentially unique at a part-of-speech level. These findings contribute to a discussion in the literature concerning the potential mitigating effects of book-reading on SES-based discrepancies in development outcomes.

keywords: Developmental psychology, early language learning, early education.

*Thanks to Drs. Dan Yurovsky and Benjamin Soltoff for their valuable input.

†jgdenby@uchicago.edu.

1 Introduction

Within the broader literature on child development, a strand of research concerns differential language development outcomes across socioeconomic strata. For decades, sociologists and psychologists noted effects of socioeconomic status (SES) on various developmental outcome measures, such as lexical maturity, grammatical maturity, and school achievement (e.g., Hoff-Ginsberg, 1998; McLoyd, 1998; Walker et al., 1994). Researchers pointed to various candidate drivers of outcome disparity (e.g., access to academic resources, cognitive stimulation at home, exposure to lead), with little consensus on the main causal difference(s) and how outcome disparity could be alleviated. However, over time, a growing focus on the linguistic environment of developing children has yielded surprising insights into why this disparity exists.

Some early research into this question of SES-based disparities probed into how caregiver speech differs between families of different socioeconomic strata. For instance, Walker et al. (1994) finds that caregivers from high SES families tend to converse more (in terms of time, attention, and word count) with their children (aged 7 to 36 months) than their lower SES counterparts, and these increases are associated with better development outcomes such as vocabulary size and school performance. Further, Hoff-Ginsberg (1991, 1998) demonstrates substantial differences in the actual content of mothers child-directed speech along the SES dimension. Specifically, mothers of high SES typically produced more speech (in raw word count), speech that was more lexically diverse (based on the number of unique word roots used), and speech that consisted of longer utterances (using the mean utterance length [MLU] measure.) These initial findings represented an early concerted effort to determine low-level differences in early linguistic input that could (begin to) explain the substantial and long-reaching discrepancies in language development across socioeconomic strata; however, since these studies were purely observational and concerned relatively small sample sizes, they lacked the power to convincingly rank caregiver speech differences among the myriad potential explanations for the impact of SES on early language development.

Later, Hoff (2003) demonstrates through a similar procedure that SES affects early vocabulary development primarily through components of maternal speech. In this study, families from two categories of SES allowed the researchers to record conversations between the mother and infant child at regular intervals. Using hierarchical statistical analyses, the researchers find that the difference in productive vocabulary growth between high-SES and mid-SES children was fully explained by differences in mothers speech, specifically differences in number of word tokens, number of word types, and sentence complexity (via MLU.) While these findings do not explicitly uncover the main mechanism underlying the effects of SES, they do serve as a beginning for that project by identifying the mediating variable. Rowe (2008) corroborates these findings with a new set of parents and further speech analysis. From this line of research, it is clear that SES-based differences in early linguistic development are in large part explicable via low-level dimensions of caregiver speech.

In a different line of research, a small area of the developmental psychology literature has investigated the role of picture books (specifically story time) in shaping a childs early linguistic environment. Around this time, several research teams (cf. Ninio, 1980; Snow and Goldfield, 1983) studied story time as a unique form of linguistic interaction between children and caregivers, finding that reading aloud often prompts instructive behaviors (such as question-asking and explicit labeling), which, in turn, correlate positively with development outcomes. Whitehurst et al. (1988) serves as one of the first prominent experiments concerning the particular content and effects of picture books on early linguistic development through parental read-alouds. In their study, they find that parents who read books according to an experimenter-prescribed strategy involving increased rates of open-ended questions, expansions upon the story, and conversational responses to their childrens interactions with the story, prompted better vocabulary development in their children when compared to the children of parents who read with a straight reading approach. These findings are interesting in that they begin to uncover the positive effects of reading on linguistic development; but, in large part, they ignore the actual content of childrens books and how it might differ from speech in a way that uniquely supplements a childs early

linguistic environment.

Recently, Montag et al. (2015, 2018) have aimed to answer this question by constructing their own corpus of popular childrens books and conducting content analyses in conjunction with a prominent corpus of child-directed speech (MacWhinney, 2000). They focus particularly on lexical diversity, using type-token ratio (a common metric in natural language processing) to assess the content of length-matched samples from each corpus. In doing so, they find that, for all sample sizes tested, text content from picture books tended to have markedly higher lexical diversity than child-directed speech. Drawing upon work discussed above, the researchers then argue that picture books, as a medium of linguistic input that tends to be more lexically diverse, may serve to uniquely benefit linguistic development through exactly this property. Moreover, as lexical diversity of mothers speech serves as a mediating variable for the effect of SES on childrens language development (Hoff, 2003), picture books may aid in relieving the challenges faced by low-SES children by expanding and enriching the dataset upon which early language learning depends.

My research closely follows the path forged by Montag et al. (2015, 2018) in both expanding their content analysis to include other dimensions highlighted by Hoff (2003) as mediators of SES effects (e.g., MLU), and replicating their findings with another prominent speech corpus consisting of a sample of families more representative of socioeconomic status (Goldin-Meadow et al., 2014). In doing so, I aim to strengthen the argument that picture books provide a unique source of linguistic input to young learners in a way that potentially mitigates the deleterious effects of low SES.

2 Corpora

My research focuses on analyzing two corpora. First, the Language Development Project (LDP) is a large-scale longitudinal effort by researchers at the University of Chicago to study child development along various dimensions (Goldin-Meadow et al., 2014). Pertinent to my research interests, it includes transcripts of parent-child conversations recorded *in situ* at regular intervals (every four months between

the ages of 14 to 58 months); I use this dataset to evaluate lingustic characteristics of child-directed speech across development. Notably, LDP consists of data from a socioeconomically diverse sample of Chicago families, making it a more robust source of data for comprehensive child development research (see Table 1 for details.)

Table 1: Summary info about LDP

Number of Parents	101
Number of Conversations	882
Prop. lowSES (<$50,000)	30%
Prop. highSES (>$50,000)	70%

I compare the speech from LDP to text from a corpus of 100 popular childern's books sourced from Montag et al. (2015). The books included in the corpus were selected by conducting surveys of parents and librarians, examining local library circulation statistics, and pulling from Amazon.com best-seller lists.

3 Analysis and Results

With this project, I am primarily interested in evaluating the content of children's book text and comparing it to that of typical child-directed speech. In order to do so, I employed computational analysis methods to investigate three linguistic dimensions: lexical diversity, sentence complexity, and part-of-speech usage.

3.1 Lexical Diversity

Following convention within the content analysis literature, I used type-token ratio (TTR) as a measure of lexical diversity. TTR serves as a measure of diversity by comparing the number of unique words to the total number of words used within a specified range. On this measure, a higher TTR corresponds to more diverse language. My analysis followed closely in the footsteps of Montag et al. (2015); I used Python scripts with the `nltk` packages to collect increasingly large samples of the corpora (with LDP split by SES) and compute unique word counts (Bird et al., 2009).

Figure 1: Lexical Diversity

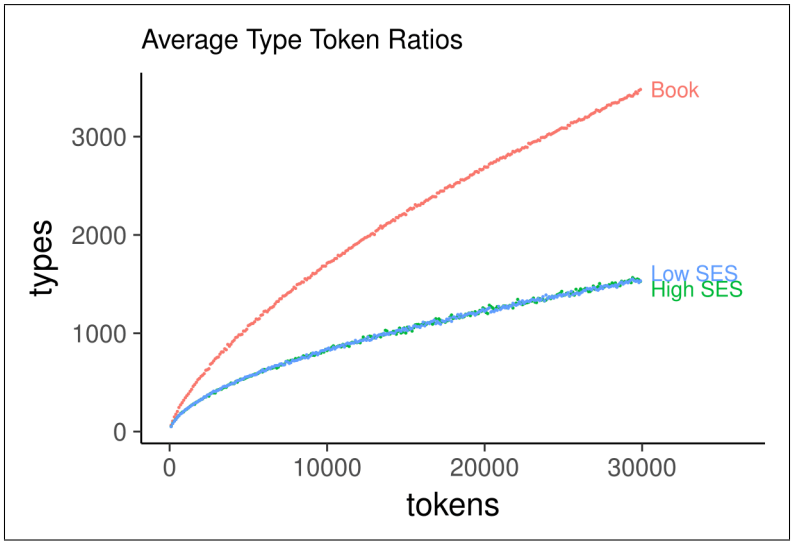
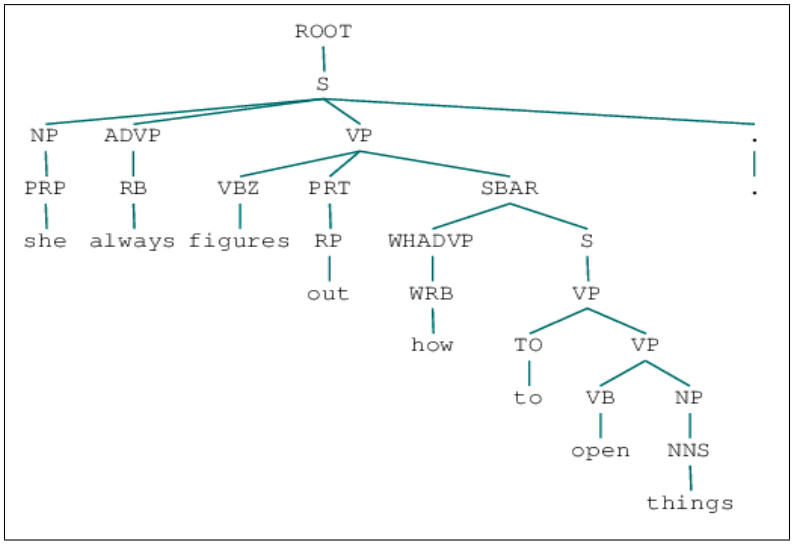


Figure 1 shows the average type count for each group (low SES speech, high SES speech, and book text) across random contiguous word samples between 0 and 30000 tokens. This plot demonstrates a strong replication of the TTR finding in Montag et al. (2015); book text tends to be much more lexically diverse than child-directed speech of any kind, particularly as the sample size grows. This finding, combined with previous findings that lexically diverse speech is positively associated with linguistic maturity, seems to lend support to the notion that book text is well-poised to supplement a child's early linguistic environment by supplying it with a richer vocabulary.

3.2 Sentence Complexity

Using the Stanford NLP group's Java implementation of a parse tree algorithm (via the `lucem_illud` Python package), I used parse tree depth as a proxy for sentence complexity. Figure 2 is an example of a parse tree with the sentence "She always figures out how to open things." Roughly, the algorithm breaks down a sentence into its component clauses and words in the order in which the sentence is syntactically composed. The method computes an internal representation of a sentence's parse tree and allows for easy extraction of its depth; for example, the tree in Figure 2

Figure 2: Example Parse Tree



has a depth (or complexity) of 10. First, I ran this algorithm on the entirety of the book corpus, computing a complexity score for every sentence in every book. Then, I randomly sampled 90 sentences (about the length of the average book) from each LDP conversation, aggregating a group of speech excerpts that were length-matched to the books. Running the algorithm on these samples, grouping by child age and SES, I created a time-series of values capturing the expected average sentence complexity spoken by a caregiver at a given point in their child's development.

Figure 3: Sentence Complexities

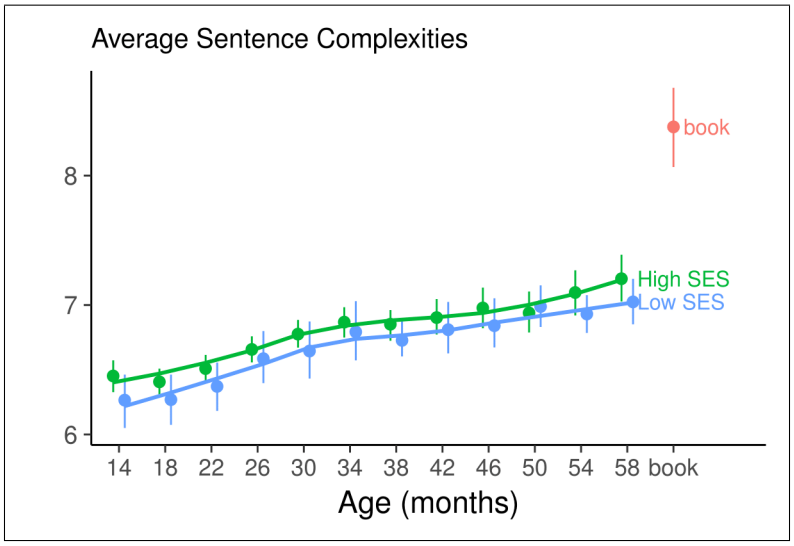
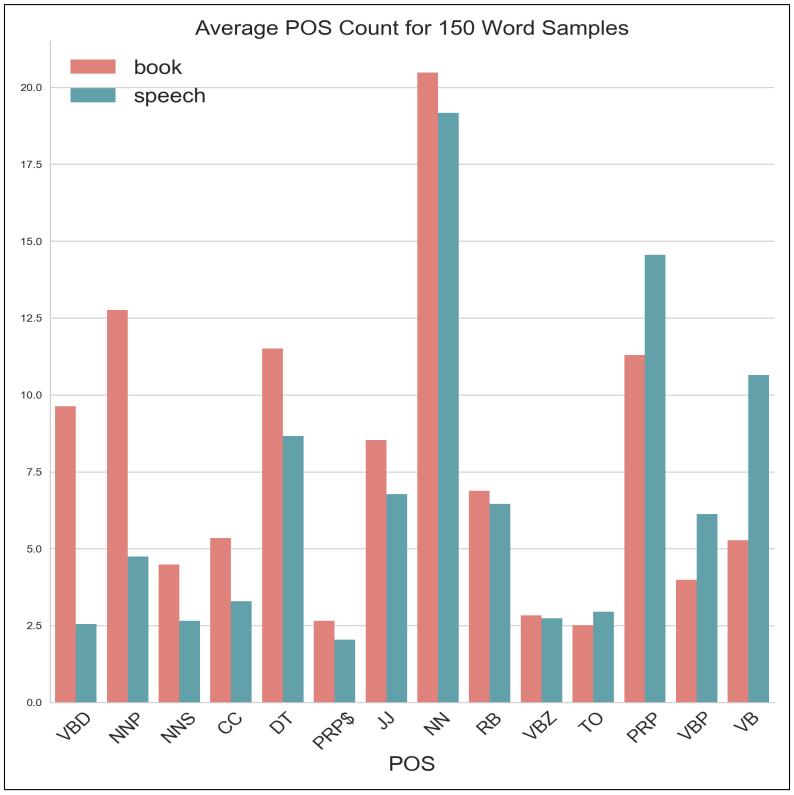


Figure 3 displays the average sentence complexity scores for the 90 sentence samples taken from parents at each time point, alongside the average complexity score computed from all sentences in the childrens book corpus (with 95% confidence intervals for every point.) Two findings stand out from this plot: First, there is a small replication of previous findings in the literature showing that parents of high SES tend to use more complex language than their low SES counterparts. This finding is not strongly significant from this analysis — this is likely due to the particular sampling strategy used. Second (and perhaps more importantly), the book corpus's average score stands far above the parental average at any time point. This illustrates that, on average, children's books contain much more syntactically complex language than parent's speech throughout early development; if less complex language in speech at a sentence level is a strong component of SES-based discrepancies in language development outcomes, here is a demonstration of how books might aid in closing that gap.

3.3 Part-of-Speech Usage

Again using the `nltk` package, I used the Penn-Treebank tagset to tag the parts-of-speech for random contiguous 150 word samples from each LDP conversation and chidrens book in the corpora. (In Figure 2, the words are also tagged by the Penn-Treebank tagset.) With this information, I could then observe average part-of-speech usage within books and child-directed speech at various levels of granularity and according to dimensions of child age and SES. Figure 4 shows the average part-of-speech (POS) usage computed over a large number of 150 word samples taken from each corpus. Most striking are the POS that differ largely between the two media. Specifically, VBD (or 'to be' in past tense) tend to be used much more often by books than by parents in typical child-directed speech, while the trend is reversed for VBP & VB (generally, 'to be' in present tense.) This suggests that books are far more likely to contribute language about the past than child-directed speech, supporting the idea that books are capable of acting as contributors of unique components of

Figure 4: Part-of-speech



language. The connection between this finding and other arguments in the literature is less clear or robust, but these differences collectively serve as a strong indicator that children's books offer a unique *kind* of language and don't just differ from speech in degree.

4 Conclusion

Many researchers within developmental psychology try to uncover the exact mechanics of language learning, and how/why language learning happens differently between children of different socioeconomic statuses. By examining speech at a very fine level, researchers have pointed to certain linguistic dimensions that might explain SES-based effects on language development. My project continues others' efforts to extend this type of analysis to another prominent domain of early linguistic input: children's books. Through computationally-enhanced linguistic analysis comparing

the text of children's books to transcripts of child-directed speech, I find that book text is noticeably different from speech along three dimensions: First, children's books tend to use a much richer vocabulary than speech (Figure 1); second, children's books have a higher average sentence complexity than speech at any time in a child's early linguistic development (Figure 3); third, children's books use past-focused language (viz., past-tense verbs) at a much greater rate than speech, indicating that book text may supply a unique *kind* of language (Figure 4). Taken altogether, my findings further the hypothesis that children's books cannot be ignored in studying inputs to a child's early linguistic environment, as their contributions are markedly different from child-directed speech. Moreover, when considering previous research concerning the role of lexical diversity and sentence complexity in explaining SES-based differences in language development, these findings may point towards a fleshing out of the mechanisms by which book-reading can help to mitigate outcome disparities.

References

Bird, Steven, Edward Loper, and Ewan Klein, "Natural language processing with Python: analyzing text with the natural language toolkit," O'Reilly Media Inc. 2009.

Goldin-Meadow, Susan, Susan C Levine, Larry V Hedges, Janellen Huttenlocher, Stephen W Raudenbush, and Steven L Small, "New evidence about language and cognitive development based on a longitudinal study: Hypotheses for intervention.," *American Psychologist*, 2014, *69* (6), 588–599.

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

Hoff-Ginsberg, Erika, "Mother-Child Conversation in Different Social Classes and Communicative Settings," *Child Development*, 1991, *62*, 782–796.

_ , "The relation of birth order and socioeconomic status to children's language experience and language development," *Applied Psycholinguistics*, 1998, *19*, 603–629.

MacWhinney, Brian, *The CHILDES Project: The database*, Psychology Press, 2000.

McLoyd, Vonnie C, "Socioeconomic disadvantage and child development.," *American Psychologist*, February 1998, *53* (2), 185–204.

Montag, Jessica L, Michael N Jones, and Linda B Smith, "The Words Children Hear," *Psychological Science*, July 2015, *26* (9), 1489–1496.

_ , _ , and _ , "Quantity and Diversity: Simulating Early Word Learning Environments.," *Cognitive Science*, February 2018, *17*, 814.

Ninio, Anat, "Picture-Book Reading in Mother-Infant Dyads Belonging to Two Subgroups in Israel," *Child Development*, June 1980, *51* (2), 587–590.

Rowe, Meredith L, "Child-directed speech: relation to socioeconomic status, knowledge of child development and child vocabulary skill," *Journal of Child Language*, 2008, *35* (01), 185–205.

Snow, Catherine E and Beverly A Goldfield, "Turn the page please: situation-specific language acquisition," *Journal of Child Language*, October 1983, *10* (3), 551–569.

Walker, Dale, Charles Greenwood, Betty Hart, and Judith Carta, "Prediction of School Outcomes Based on Early Language Production and Socioeconomic Factors," *Children and Poverty*, April 1994, *65* (2), 606–621.

Whitehurst, G J, F L Falco, C J Lonigan, J E Fischel, B D DeBaryshe, M C Valdez-Menchaca, and M Caulfield, "Accelerating language development through picture book reading.," *Developmental Psychology*, 1988, *24* (4), 552–559.