Methods & Results

Joseph Denby

5/9/2018

Data Sources/Collection

My project uses two prominent corpora as its primary sources of data. First, the Language Development Project (LDP) consists of transcribed conversations between parents and children. More specifically, the dataset includes text from parent-child dyads between the ages of 14-58 months, with data collected at four-month intervals. The dataset also includes a separate demographics document pairing each subject's anonymous ID number with their personal information (e.g., income, education, race, etc.) As the income information was relevant to my analysis, I extracted this variable from the demographics data and paired the data with the conversation dataset. Access to the dataset was obtained by contacting the researchers involved in its organization and fulfilling the eligibility requirements for researcher access. Below is a table with summary information about the LDP dataset:

LDP Summary Table	
Number of Parents	101
Number of Conversations	882
Prop. earning less than $50,000 (lowSES)	30%
Prop. earning more than $50,000 (highSES)	70%

Next, I used the Children's Book Corpus (Montag et al., 2015), a collection of 100 popular children's books as recommended by parent & librarian surveys, Amazon.com best-seller lists, and library circulation statistics. By contacting the corpus's authors, I obtained a raw text file with the corpus contents; after some brief pre-processing and cleaning, the dataset was ready for analysis.

Procedure

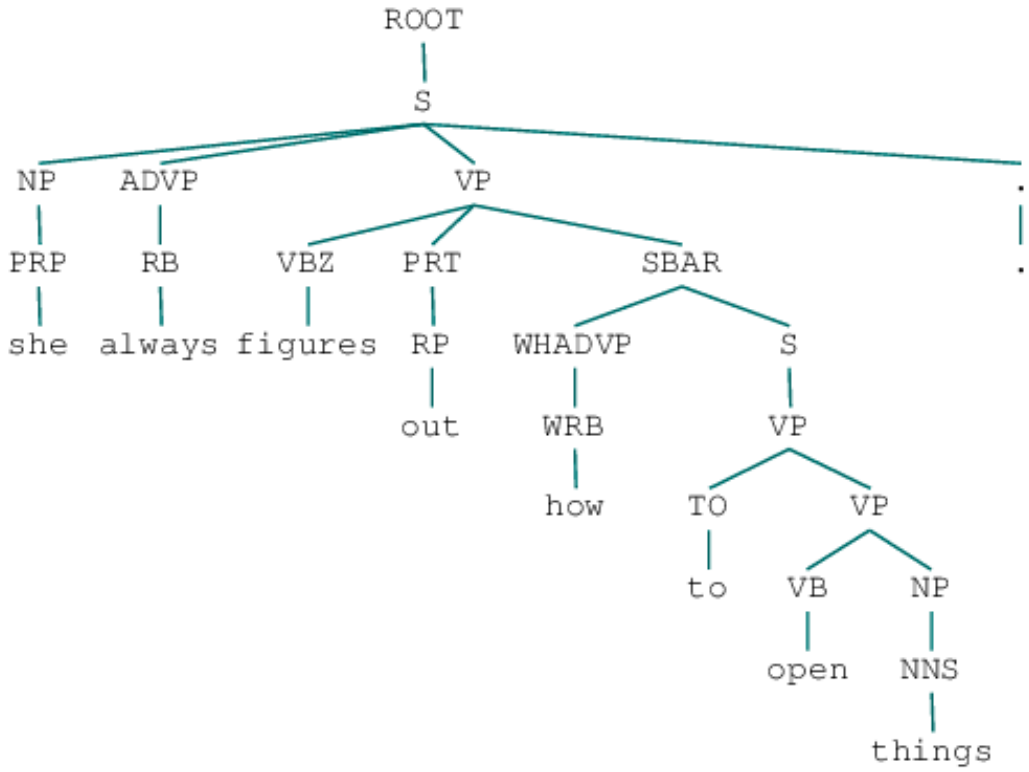
With this project, I am primarily interested in evaluating the content of children's book text and comparing with that of child-directed speech. In order to do so, I employed computational analysis methods to investigate three linguistic dimensions: lexical diversity, sentence complexity, and part-of-speech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

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 nltk to collect increasingly large samples of the corpora (with LDP split by SES) and compute unique word counts.

Sentence Complexity

With an interface to the Stanford NLP group's Java implementation of a parse tree algorithm contained within the lucem_illud Python package, I used parse tree depth as a proxy for sentence complexity. Printed below is an example of a parse tree:



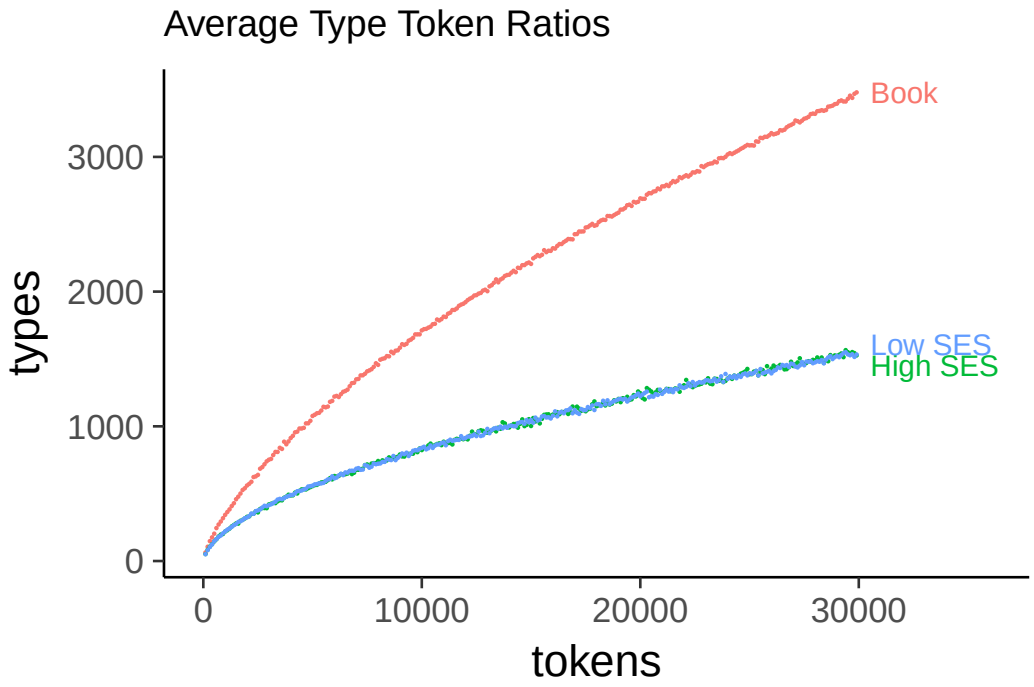
The method computes an internal representation of a sentence's parse tree and allows for easy extraction of its depth; for example, the above tree has a depth (or complexity) of 10. First, I ran this algorithm on the entirety of the book corpus, computing a complexity score for every sentence in every book. Then, since an average LDP conversation is 10 times longer than the average book, I randomly sampled 90 sentences (about the length of the average book) from each LDP conversation. Running the algorithm on these samples, grouping by child age and SES, I created a distribution of values meant to capture the expected range of sentence complexities spoken by a caregiver at a given point in their child's life.

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 chidren's book in the corpora. (In the parse tree above, 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.

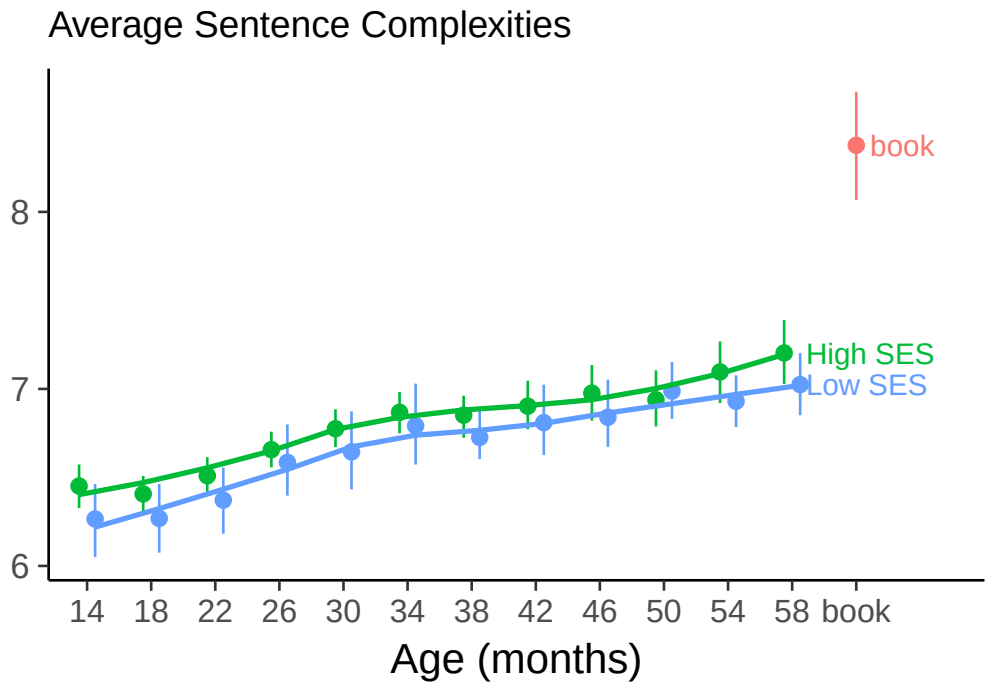
Initial Results

Lexical Diversity



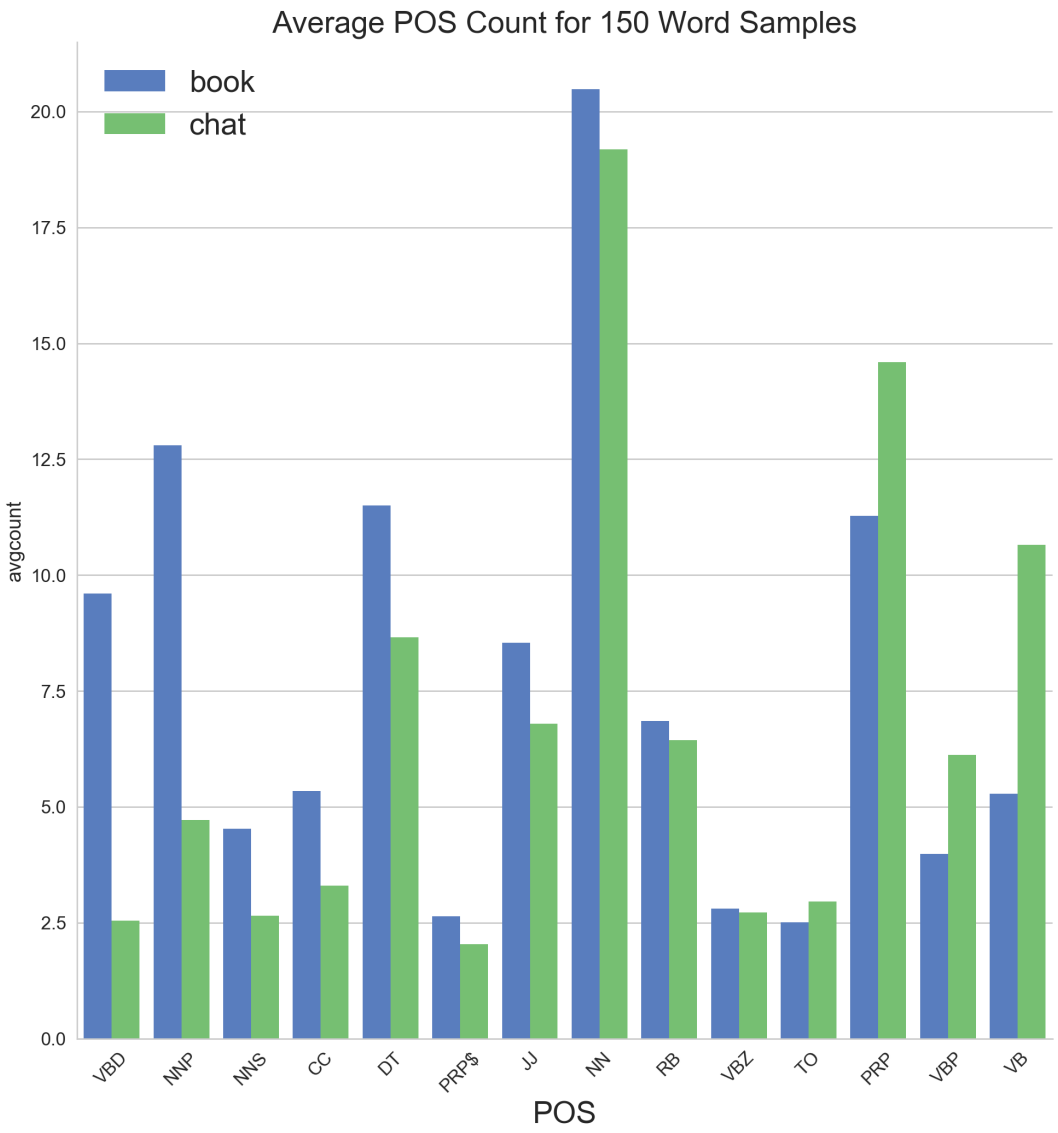
The plot above shows the average type count for each group (low SES speech, high SES speech, and book text) across random word samples between 0 and 30000 tokens. This plot's findings is a strong replication of the TTR finding in Montag et al., (2015); book text tends to be much more lexically diverse 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.

Sentence Complexity



This plot displays the distribution of average sentence complexity scores for the 90 sentence sample taken from each parent at each time point, alongside the average complexity score computed from all sentences in the children's book corpus (with 95% confidence intervals for every point.) Two findings stand out from this demonstration: 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, if reading a random sentence from a random children's book, one would expect a much more complex language than one would expect to hear from a parent when speaking to their linguistically-developing child. Again, this finding is supportive of the notion that children books are well-poised to supplement an early linguistic environment with more complex language.

Part-of-Speech Usage



Finally, the above plot 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 seems to suggest 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 language.