Joseph Denby

Professor James Evans

Computational Content Analysis

March 16, 2018

Within the literature on early linguistic development, a prominent topic concerns the substance and role of early language inputs. While historically scholarship has focused almost exclusively on caregiver child-directed speech, recent work has highlighted children’s picture books as a potentially significant source of early linguistic exposure. In order to further probe the content of children’s picture books and directly compare it with caregiver child-directed speech, I used computational analytical techniques on two relevant corpora: the Language Development Project (Goldin-Meadow et al., 2014) and 100 Children’s Books Corpus (Montag et al., 2015.) By investigating and comparing relative part-of-speech (POS) usage as well as sentence complexity, I provide inroads to a more robust understanding of the structural content of children’s books, hinting at their unique contributions to a child’s early linguistic environment. In what follows, I will provide a detailed report of the techniques used, as well as a discussion of results and their significance. But first, I will offer a more comprehensive picture of the corpora used.

**Corpora**

As mentioned above, the two corpora used for this project were selections from the Language Development Project (LDP) and a corpus of popular children’s picture books sourced from Montag et al., 2015. LDP is an initiative (based at the University of Chicago) tasked with documenting the linguistic development of children selected from a socioeconomically diverse sample of Chicago-based families. Data collection involved recording conversations (approximately 1.5 hours) between children and caregivers once every four months between the ages of 14 and 58 months. I gained access to the LDP database through my advisor, Dr. Daniel Yurovsky, who has several projects affiliated with the project. Through him, I was able to extract the information required for my purposes; after accounting for sessions with insufficient or missing data, transcripts for 885 conversations between 102 parent-child dyads were included in analysis.

The second corpus included in analysis consisted of the text from 100 popular children’s picture books. As described in the source publication, books were selected for inclusion based on librarian recommendations, best seller lists from amazon.com, and circulation statistics from the researchers’ local library. As such, the books contained within the corpus are taken to be a representative sample of what book American children are typically exposed to. I obtained the corpus by contacting the publication’s lead author, who sent me a raw text file. After some brief extraction and cleaning, I converted the corpus into a usable format primed for python-based analysis.

**Analysis**

**Preparation**

Before anything else, I prepared the text data in various ways to facilitate the particular analysis I wished to conduct.

**POS Tagging**

First, I