Chloe Zehr

CRDDS Data Bootcamp Micro-Credential

29 June 2023

**Question:**

What does POS tagging reveal about noun co-occurrence in interviews with formerly enslaved people from the early twentieth century? How might noun co-occurrences reflect historical consensus and do patterns emerge?

**Dataset, variables, and analysis:**

\*\*\*Disclaimer – the data used in this project contains racist profanity due to its context and content, I have done my best to not reinforce, nor use any of this language within my analyses\*\*\*

For this project, I generated a text corpus comprised of the first four volumes of South Carolina’s *Slave Narratives: A Folk History of Slavery in the United States from Interviews*, accessible through the Project Gutenberg online. Using the “gutenbergr” package in RStudio, along with the “tidyverse” and “udpipe” packages, I downloaded all four volumes of text and basic metadata (Gutenberg ID) directly into my environment. The combined text for all four volumes is stored in a dataframe/variable called “scnarratives\_works.”   
 Using part of speech tagging with the package “udpipe” I created the dataframe/variable “scnarratives\_annotated,” which stores all the POS tags for the corpus. Using this variable, I generated noun co-occurrence data to create a network graph of terms used in the corpus. Because my corpus is small, I used a skipgram model to iterate over the text for co-occurrence within a 30-term window (which is standard for basic textual analysis). Then I used the occurrence function to determine noun cooccurrence and produce a network graph to visualize the nodes and edges between frequent terms.

The network graph which is included in the GitHub repository highlights three significant findings. First, the visualization makes clear the semantic connection between terms like “master” and “slave” from the perspective of formerly enslaved interviewees. In contrast to colonial archives, where terms, such as “slave,” are more associated with commodities, labor, and trade, the dataset from South Carolina interviews highlights alternate perspectives (though in many ways it still reflects the systemic oppression of white supremacy and racial slavery). For instance, the term slave also co-occurs with the word “mother,” suggesting that, unlike contemporary white authored sources, interviewees understood their past status as enslaved and yet maintained their identities as mothers, families, and communities. While this all may be obvious, it is important to note that datasets such as this highlight that it is possible, and imperative, to use digital methods (and all tools available to us) to continue to question the hegemony of traditional archives and to help recover subaltern perspectives within their own contexts.

My dataset comes from the Project Gutenberg repository which is a free and open-source resource – thus my dataset does not require a license and is easily accessible to those who have access to the web. Gutenberg offers multiple formats for reading eBooks as well as generating audiobooks. In addition, Gutenberg is compatible with most dyslexic font browser extensions, which helped me greatly when becoming more familiar with my data (i.e., close reading) before performing my computations. For working with text data, Gutenberg meets FAIR’s principles of being findable (within the limits of what Gutenberg has), accessible, interoperable, and reusable. I believe that Python users can also access Gutenberg, and the online resource allows all users to download works as html files, text files, and eBooks that are compatible with devices like kindles.

However, the gutenbergr package does make using the corpus I created significantly easier. Thus, those who do not have access to R or RStudio (both are free) for several reasons, may find it difficult download and explore the metadata made available through the package’s functions.

For this dataset, one of the most challenging aspects is that packages such as udpipe as well as data such as R’s “stop\_words” rely on heavily on modern English language. The South Carolina interviews with ex-enslaved people replicate individuals’ dialects as well as local early twentieth century English characteristics – therefore, it is likely that there is a higher degree of error in computing semantic relationships and unfolding analyses. This also means that there is inconsistent vocabulary and spellings used throughout the corpus which require more time and consideration to evaluate.

A special thanks to the team at CU Boulder’s CRDDS for the excellent help and guidance.