**Data Appendix**

Jones, Jennifer J. 2016. “‘Talk like a man.’ The Linguistic Styles of Hillary Clinton, 1992–2013.” *Perspectives on Politics* 14(3).

**I. Summary statistics for the Hillary Clinton transcript corpus (N=567)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Statistic | Mean | St. Dev. | Min | Max |
| Word Count | 1,916.82 | 1,731.50 | 304 | 18,860 |
| Pronouns | 17.05 | 2.20 | 8.83 | 24.35 |
| First-person singular | 3.35 | 1.84 | 0.34 | 11.04 |
| Verbs | 17.28 | 1.79 | 11.24 | 23.02 |
| Auxiliary verbs | 11.14 | 1.26 | 6.00 | 15.26 |
| Social references | 11.05 | 1.78 | 5.75 | 20.87 |
| Positive emotion | 3.86 | 0.89 | 1.43 | 7.99 |
| Negative emotion | 1.30 | 0.63 | 0.00 | 4.47 |
| Cognitive mechanisms | 20.02 | 1.82 | 11.61 | 26.22 |
| Tentative words | 2.42 | 0.80 | 0.22 | 4.90 |
| Words >6 letters | 18.42 | 2.60 | 11.14 | 29.94 |
| First-person plural | 3.19 | 1.29 | 0.30 | 8.65 |
| Articles | 6.77 | 1.15 | 2.71 | 11.12 |
| Prepositions | 13.79 | 1.17 | 9.68 | 17.42 |
| Anger words | 0.42 | 0.33 | 0.00 | 1.78 |
| Swear words | 0.004 | 0.02 | 0.00 | 0.29 |
| Feminine/Masculine ratio | 2.09 | 0.37 | 1.11 | 3.36 |

**II. Procedures for Processing and Preparing the Transcripts for Analysis**

A number of interview transcripts were downloaded as PDFs from the Clinton Digital Library, part of the Bill Clinton Presidential Library. The text on these PDFs, however, was not machine readable—transcripts were captured as an “image’ of a printed transcript. In order to make the text machine readable, I processed the PDFs through optical character recognition (OCR) (Adobe Acrobat and Google Docs both have this functionality), and then extracted the PDF text to a text file.

OCR is not perfect, and it often inserted noisy elements, especially at the beginning of a document (e.g. sequences of periods, exclamation marks, tildes, etc.). This was especially true for poor quality and poorly scanned transcripts. Despite this, most transcripts were still readable and I was able to process the noisy elements through regular expression (discussed below). However, there were some transcripts for which OCR extracted wholly unreadable text and these were omitted from the analysis. This explains why there are more interview transcripts in the Clinton Digital Library than are included in this analysis. Note that the Cambridge archive accurately reflects the transcripts used in the analysis. All OCR text files were manually inspected for accuracy.

Once I had a corpus of text files, I catalogued each transcript’s metadata (interview source/interviewer and date) and then removed all metadata from the text files. At this point, any duplicated transcripts were removed. I then processed all text files to ensure OCR noise, questions posed by the interviewer(s)/moderator(s), comments by speakers other than Clinton, and transcriber notes— e.g. [INAUDIBLE] or (laughs)— were removed. To do this I used TextWrangler, a text editing program developed by Bare Bones Software, because it allows for multi-file editing and is capable of regular expression (RegEx) searching. Using RegEx searching and replacing, I “cleaned” irregular text and file structures (e.g. whitespace before and after the text, non-ASCII and control characters) and removed all text that was not Clinton’s language. TextWrangler has a convenient interface which allowed me to visually inspect the transcripts as I processed them.

This was a tedious process, to say the least, but it ensured that only Clinton’s language was included in the analysis.

Once the corpus was processed to include only Clinton’s language, I verified that each transcript had a minimum of 300 words, a commonly recommended cutoff for the type of analysis performed. Any transcript where Clinton spoke fewer that 300 words was omitted. I then prepared the transcripts for processing through Linguistic Inquiry and Word Count (LIWC), a text analysis program developed by Pennebaker, Booth, and Francis (2007). As a computer program, several steps needed to be taken to ensure that LIWC accurately categorized Clinton’s language:

In order to have an accurate count of words per sentence, transcripts were processed to ensure that end-of-sentence markers were indeed end-of-sentence markers. Common abbreviations (such as “Dr.”, “Ms.”, “U.S.”) were processed to remove end-of-sentence markers. For example, “Ms.” was changed to “Miss” and “U.S.” was changed to “USA”.

Transcripts were also processed to ensure that LIWC accurately captured meaningless fillers and nonfluencies, such as you know and well. After inspecting the context of these fillers, I changed all instances of you know, (with a comma) and Well, (with a comma and capital “W”) to youknow and rrWell respectively. This is in accordance with the LIWC2007 operating guidelines.

Finally, I processed the text files through LIWC, which generated the output seen in the file HRC\_LIWC\_output.csv (available on the Cambridge Digital Archive). I conducted all analysis on the LIWC output using R. This code is also available on the Cambridge Digital Archive.

**III. Transcript Sources**

**Interviews**

Clinton Digital Library, First Lady's Office, Press Office, Lissa Muscatine, Collection no. 2011-0415-S, William J. Clinton Presidential Library, <http://clinton.presidentiallibraries.us/collections/show/15>, accessed September 2014.

The US Department of State, Former Secretary Clinton's Remarks [Interviews Only], <http://www.state.gov/secretary/20092013clinton/rm/index.htm>, accessed October 2014.

The US Department of State, Former Secretary Clinton's Town Halls and Townterviews [Interviews, Town Halls and Townterviews only], http://www.state.gov/secretary/20092013clinton/townhalls/index.htm, accessed October 2014.

Digital archives: LexisNexis, ProQuest, Factiva, EBSCO Host/Academic Search Premier

**Debates**

The American Presidency Project,<http://www.presidency.ucsb.edu/debates.php>*,* compiled by Gerhard Peters and John T. Woolley. Accessed October 2014.

**IV. Notes on Transcripts Uploaded to the Cambridge Digital Archive**

Transcripts uploaded to the Digital Archive are pre-processed, fully intact interviews and debates—with transcript metadata, interviewer questions, etc. Unlike the raw transcripts downloaded from the Clinton Library, however, the PDFs uploaded to the archive have been digitized through optical character recognition (OCR). This means the transcripts are now searchable and the text within the PDFs can easily be extracted to a text file (Adobe Acrobat and Google Docs both have this functionality), which is what I did for this analysis.

Transcripts sourced from the Clinton Library are formatted as OCR-optimized PDFs. Transcripts sourced from the U.S. State Department are organized by the year the interview took place and formatted as a single tab-separated value (.tsv) file for each year. Transcripts sourced from Lexis Nexis, Factiva, or EBSCOhost are formatted as text files. Note that in a few cases (post 2000) multiple interviews are contained on a single text file. These are labeled accordingly.