ENGL 1400 Introduction to Literary Studies
Mary Loeffelholz
Corpus-building Handout for Computational Text Analysis

Taking it Further with Corpus Building

In this session, we used web-based text analysis tools with a pre-built corpus, but there is a lot more you can do! In this handout, you will find some more information on how you could build and analyze corpus of your own.

Key Words

- **Computational Text Analysis**: Text analysis is making inferences based on textual data. Computational text analysis (CTA) involves a computer drawing out patterns in a text, and a researcher interpreting those patterns. CTA includes methods such as word count frequency, nGrams, and sentiment analysis. CTA is similar to statistical analysis, but the data are texts.
- **Corpus (plural–corpora)**: A collection of texts used for analysis and research purposes.
- **Stop words**: words that appear frequently in a language, like pronouns, prepositions, and basic verbs. These are often removed for computational analysis. Some English stopwords include: a, the, she, he, I, me, us, of, is, would, could, should, etc.
- **Word Count Frequency:** Counting the total times a word appears in a text/corpus or the percentage of how often it appears.
- **Sentiment Analysis**: Measuring the sentiment of a text based on a scale such as negative/positive or happy/sad. Each word has a particular weight to determine where on the scale it falls, and these weights are calculated to determine a text's overall sentiment.

How to Build a Corpus

When building a corpus, especially in the context of a smaller project, follow these steps:

1. Choose the texts you would like to include in your corpus.
    - Remember, these texts are not necessarily representative of a larger body of writing, and that the texts you select will have a significant impact on your results.
    - In any argument and analysis of your results, you should specifically address and analyze the contexts of these texts and consider any possible limitations in their ability to serve as proxies for the phenomena you wish to study.
2. Create a folder on your computer or cloud storage where you will store your corpus. You might title this folder "browne_corpus".
3. Once you have chosen the texts you will include, open a plain text editor (for example, Notepad on PCs and TextEdit on Macs).
    - TextEdit on Macs: You must make sure it is configured to work with plain text files. To do this, open Text Edit and go to "Preferences" and make sure "plain text editor" is selected. Then, restart TextEdit.
4. Making plain-text files:

Developed by: Sarah Connell, DITI Co-Director
Colleen Nugent, Margarida Rodrigues, and Yunus Emre Tapan, DITI Fellows
Questions? Contact us: nulab@northeastern.edu

- The individual plain text (.txt) files that make up your corpus are stripped-down and machine readable versions of the documents (PDFs, .doc, .docx, etc.) you chose to include in your corpus.
- To add the actual text, you would simply copy and paste the contents of each document into the text editor. Step-by-step instructions for this process are included in the slides.
  - Often, texts that are on websites can easily be copied and pasted; however, copying and pasting the text from documents can take a bit more time and require more extensive data cleaning.
  - **Only copy one text** into each new plain text file (unless you are combining texts from similar resources for research purposes).
- Some articles might have HTML/web-browser versions that will be easier to copy-and-paste than PDFs.
- If you cannot copy and paste the text (if it is a PDF or an image), either find a text that you can copy and paste, or transcribe the text.
- Make sure each file name ends with .txt – this is a plain text file and most GUI tools will accept these.
- Use filenames to indicate the data inside (ex: "browne_chapter-1.txt")
- Make sure not to put any spaces in the names of the files as you save them. Use underscores or hyphens to mark spaces between words instead.

5. Repeat steps 4 and 5 for each text in your corpus.
   - For example, if you have five texts in your corpus, create five files.

Where to Find Texts for Corpora

You can start by browsing NULab's datasets, which contains links to repositories of text, including the Gutenberg project - a repository of over 60,000 free eBooks in plain text - or Northeastern's Early Caribbean Digital Archive (ecda), containing pre-twentieth century Caribbean texts, maps, and images.

Web-Browser Computational Text Analysis Tools

These browser-based GUI (Graphical User Interface) text analysis tools can show word frequencies and patterns in language. While using coding languages like Python and R can open up other types of analysis (such as word embedding models and topic modeling), these GUI tools allow you to do more basic analysis to begin examining your texts computationally. We will be working with:

- **Word Trees**: This is a good way to see patterns in word usage, based on words that appear before and after a term or terms of interest. There are some restrictions in size; fewer than 1 million words should work, but loading that much text in might be slow. http://jasondavies.com/wordtree/

Developed by: Sarah Connell, DITI Co-Director
Colleen Nugent, Margarida Rodrigues, and Yunus Emre Tapan, DITI Fellows
Questions? Contact us: nulab@northeastern.edu

- **Word Counter**: This is a user-friendly basic word counting tool; it allows you to count single words, bigrams, and trigrams in plain text files and to download spreadsheets with your results. The max file upload is 10MB. http://databasic.io/en/wordcounter/
- **Lexos**: This is an excellent tool for preparing and analyzing digital texts; it offers several options for text preparation, and a wide range of different analytical possibilities as well. Importantly, it also preserves all the changes that are made to a text, so that any results can be reproduced. http://lexos.wheatoncollege.edu/
- **Voyant**:This suite of tools gives you counts of words and lets you compare patterns in word locations and frequencies, or examine keywords in context, along with a few other options. Voyant will let you upload larger files than most other interfaces (up to as many as 4 million words, though it may take more than one try to successfully upload very large files). http://voyant-tools.org/

Other Computational Text Analysis Tools

Beyond the web-based tools used in our session, there are more advanced tools that you can explore. AntConc, described below, adds a layer of complexity to web-based tools but doesn't require any command-line coding.

- **AntConc**: A corpus analysis toolkit for concordancing and text analysis. AntConc is free and available to download for MacOS and Windows. AntConc performs better with many small files, rather than one or two large ones—there is no limit on how many words you can analyze, but larger corpora will take longer to work with https://www.laurenceanthony.net/software/antconc/. You can find a useful tutorial on AntConc by the Programming Historian, where you can also find a range of other tutorials on analyzing texts. Another helpful AntConc tutorial is made available by Manhattan College.
- If you are looking for more advanced tools and lessons, check out the NuLab text analysis resources and this handout.

Developed by: Sarah Connell, DITI Co-Director
Colleen Nugent, Margarida Rodrigues, and Yunus Emre Tapan, DITI Fellows
Questions? Contact us: nulab@northeastern.edu