ENGL 3370: How to Build a Corpus

Tools & Methods

May 22, 2019

Cara Marta Messina
Molly Nebiolo

Northeastern University
NULab for Texts, Maps, and Networks

Feel free to ask questions at any point during the presentation!

Workshop Learning Objective

What should we consider when creating a corpus, or a collection of texts, and how do we build one?

Example

https://drive.google.com/drive/folders/1qUnsbKe578fTSGtZMsr0cI7LI8Q2NJN

What questions might have gone into the creation of this corpus? How might this set of texts answer those questions?

Pre-Research Research - It's a thing.

Thinking and questioning what you want out of a corpus is part of the "pre-research research" that is needed to build a corpus

Method: Computational Text Analysis

One method you will be using for this research project **computational text analysis** (this method can be found in disciplines such as writing analytics, digital humanities, writing studies research, literary criticism, fields in the social sciences, and more).

In order to use this method, you will need to:

- Be familiar with the archive or the space from which you're collecting your data
- Choose an array of texts that may fit your research question, or a random selection of texts for a more unbiased result

Corpus Building: Text Selection

We recommend to choose at least **fifteen** texts from the archives

- You may choose texts that **relate directly** to your research question. This will help you tackle questions such as “how do incarcerated writers describe their experiences with X?”
- You may also choose **a random assortment** of texts as a more exploratory and less biased process. This will help you tackle questions such as “What are the linguistic patterns in X archive? What do these patterns suggest about X?”

Corpus Building: Storing Your Texts

Once you have chosen the texts you want to include in your corpus, you must **store** these texts. You can get creative with your storage, although we recommend being systematic, but this is one basic way:

1. Create a **folder** titled “corpus”
2. Save individual texts as **.txt** files (using TextEdit, NotePad or any **plain-text** editor) in the folder; .txt files are best for computational text analysis. You can **copy and paste** the texts from the site into the .txt files (or transcribe if there is no copy-pasting available, like with photos of letters)
3. Follow a **naming convention**, or a practice in which you name your files to make them distinguishable and easy to navigate. Do **not use spaces in file names**.
4. Save metadata. You can do this in a separate .txt file (as we did), use a spreadsheet, or any other method for easy metadata information storage. Check out “zz_metadata.txt” as an example.

TL;DR…… Things to Remember!

- Your results are a **reflection of the data you choose** to analyze.
  - They are *not* generalizable
  - Make your choices transparent
- **Read** and **reflect on** the texts you put in your corpus!
- **Computational text analysis** results (next week's class) may represent the data differently than **your experience** while reading and reflecting on the texts.

Thank you!

If you have any questions, contact us at:

Cara Marta Messina
Digital Teaching Integration
Assistant Director
messina.c@husky.neu.edu

Molly Nebiolo
Digital Teaching Integration
Research Fellow
nebiolo.m@husky.neu.edu

Northeastern University
NULab for Texts, Maps, and Networks