

**A Historical Sociolinguist's Digital Tools Starter Kit
[a hands-on workshop]**

Kelly E. Wright
University of Kentucky

Conference: North American Research Network in Historical Sociolinguistics (NARNiHS)
21-23 July 2017 – University of Kentucky

This workshop will present a baseline of Open Source digital tools for historical sociolinguistic investigation. I believe this will be possible, and most beneficial, if we are all working with a prepared dataset. The Parsed Corpus of Early English Correspondence (PCEEC) provides a rich, accessible data set for the kind of hands-on introductory workshop proposed here. I have prepared manageably sized and uniformly formatted sections of this dataset, which can be access here: <http://www.uky.edu/~mrlaue2/narnihs2017/workshop.html>.

We will explore various angles of inquiry by using the following tools on the PCEEC, providing a more faceted appreciation for the data. The workshop will begin with a Regular Expressions (RegEx) primer/refreshers to enhance use of basic search strings in a text/code editor (e.g. BBEdit or Notepad++) or in the corpus analysis software, AntConc. This will be followed with a short introduction to the spelling normalization software VARD2. Following this, the group will experiment with social network visualizations using Gephi. If time allows, we will dig deeper into the Associated Information File, which contains various demographic information on the authors and readers of the PCEEC letters, and discuss approaches to documentation, preservation, and standardization of metadata.

These activities will illustrate that with some foreknowledge of a dataset, one can maximize search coverage and minimize noise in the results by using the special set of characters RegEx provides to help define search patterns. Changes/Variation in spelling and the estimations of naïve writers can thwart the accuracy of digital searches across a body of historical texts. VARD2 allows for variant spellings in the data to be gathered into a search using the current (or contemporary) standard. Using PCEEC node and edge lists, we will use Gephi to view several social networks, revealing the capacity for quantification and qualification of sociolinguistic interaction that such analysis adds to historical data.

Along with the PCEEC selection, I have made available to NARNiHS access to a full set of guided instructions, complete with screenshots. This will be useful for members who cannot attend the workshop or stay for its full duration, but are interested in the tools.