Assignment 6
L615
Due Monday, April 22

1. Using the BooTCaT tools, create 2 small comparable corpora from the web, which are in 2 of your favorite languages (or your 2 favorite regional dialects). In other words, create corpora for the same topic in 2 different languages, and when I say “small”, that means you can use the default settings (e.g., 10 URLs per query).
   (a) Report exactly what you did, including, e.g., the seed terms for each language.
   (b) Report the following basic statistics for each corpus:
      - Number of tokens
      - Number of types
      - Number of URLs
      - Average document length (in words)
   (c) Compare the 20 most frequent words in each corpus. Do you notice any particular trends?

2. I’m not going to give you an additional assignment for your final project, but will just outline the final details for what you need to do.
   (a) On either Monday, April 22 or Wednesday, April 24, you will give about a 7–10 minute presentation to the class on your corpus project. This is less something that you’ll be graded on and more a chance to share work with each other and receive feedback from me. On the 15th, I’ll bring a sign-up sheet to class, so you can sign up for a slot.
   (b) Your final paper (10-20 pages) will be due on Monday, April 29 at 5pm.

If you have any questions about any components of the project, let me know.