Towards Large-Scale Computational Dialectology and Sociolinguistics

Sociolinguistics and dialectology study how language varies across socially-distinct groups of speakers. While these fields feature a strong quantitative tradition, the standard methodology requires the researcher to specify the linguistic dimensions of variability in advance—before correlating them against extra-linguistic factors. Moreover, much of this work depends on interviews for gathering data, raising problematic issues of how to elicit vernacular speech in an interview setting. However, the rapid growth of social media offers exciting new possibilities for the study of socially-oriented linguistic variation.

Using a new corpus of geo-tagged text from Twitter, we have developed new computational techniques for studying linguistic variation from raw text. These methods are capable of identifying both coherent linguistic communities as well as specific lexical features that distinguish social and geographical groups. Applying these methods to Twitter, we have discovered new and robust lexical-geographic relationships that were undocumented in prior work. In addition, we are able to accurately predict microblog authors’ geographic location from their raw text.

Jacob Eisenstein is a postdoctoral fellow in the Machine Learning Department at Carnegie Mellon University. His research focuses on machine learning for discourse, non-verbal communication, and social media. Jacob completed his Ph.D. at MIT in 2008, winning the George M. Sprowls award for his dissertation, “Gesture in Automatic Discourse Processing.”

http://www.cs.cmu.edu/~nlp-lunch/