Text Mining for Sociolinguistic Research

Joseph Roy
Anna María Escobar
Kate Lyons
Gyula Zsombok
(all organizers University of Illinois at Urbana-Champaign)

This workshop will illustrate how to use text mining and more specifically, topic modeling (Blei, 2012) of a transcribed corpus in order to model linguistic variation and change and will demonstrate the results of such analysis on three different sets of data. The process involves taking a corpus of text and reducing every text, for example sociolinguistic interviews, to a set of topics with associated probabilities for each topic.

The extraction and analysis of topics can have several benefits for sociolinguistic research in that it will allow researchers to incorporate "topic" into regression models as a gradient variable. Speaker attitudes can be extracted from open-ended questions rather than questions that require a closed answer (e.g., Ethnic Orientation in Hoffman & Walker, 2010) using the same multidimensional reduction analyses. We present three applications of this technique in three different languages and across a variety of data sources: sociolinguistic interviews, formal news articles and social media data.

Outline of Workshop
- Introduction
- Preparing the data
- Topic modeling
- Using topic probabilities in a regression
- Application 1: Effects of Contact in Modern Andean Spanish (Escobar & Roy)
- Application 2: Lexical borrowing in French print media (Zsombok)
- Application 3: Linguistic landscapes on Social Media (Lyons)
- Conclusion & Discussion

The R code used to generate results in the workshop will be made available to registrants via Dropbox, but the workshop is not hands-on (so you do not need to install anything prior or bring a laptop): it is a discussion of the methodology, choices researchers will face when using these techniques, and applications.

References

Blei, D. M. 2012. "Probabilistic topic models." *Communications of the ACM* 55, 4: 77-84.
Hoffman, M. F., & J.A. Walker. 2010. Ethnolects and the city: Ethnic orientation and linguistic variation in Toronto English. *Language Variation and Change*, 22, 1: 37-67.