Big Data and Sociolinguistics

Josef Fruehwald
University of Edinburgh

Abstract: Along with many other strains of evidence, sociolinguistics of the future is going to be utilizing "Big Data," as are most social sciences. Of course, this means that sociolinguistic researchers are going to need to get training in building and using tools for managing and analyzing our growing data sets. But more importantly, we will have to exercise *caution,* in our analyses, since non-zero effects are more achievable with larger data sets, and we will have to continue to be more serious about theory. A fundamental challenge in the social sciences is to develop theories that make precise quantitative predictions. In the spirit of the futrospective nature of the panel, I'll quote Guy (1991): "The development of models that have explanatory value in this sense - models from which one can derive precise quantitative predictions - is one of the fundamental challenges facing our discipline." In this talk I'll walk through two examples of this approach to quantitative prediction based on data from the Philadelphia Neighborhood Corpus. First, I'll look at the effect coarticulatory forces are predicted to have in circumscribing a phonetic change in progress. Second, I'll estimate the possible socio-indexical information conveyed about a speaker's gender by their filled pause preference, utilizing basic information theory. Neither of these analyses would have been possible without larger data sets than sociolinguists have previously worked with. Both require some theoretical precision (the data does not just talk for itself), and both sets of results run counter to conventional thinking on these matters.