Sociolinguistics for Computational Social Science

Sali Tagliamonte
University of Toronto

Abstract

In recent years, a major growth area in applied natural language processing has been the application of automated techniques to massive datasets in order to answer questions about society, and by extension people. Sociolinguistics, which combines anthropology, statistics and linguistics (e.g. Labov 1994, 2001), studies linguistic data in order to answer key questions about the relationship of language and society. Sociolinguists focus on frequency and patterns in linguistic usage, correlations, strength of factors and significance, which together reveal information about the sex, age, education and occupation of speakers/writers but also their history, culture, place of residence, social relationships and affiliations. The findings arising from this type research offer important insights into the nature of human organizations at the global, national or community level. They also reveal connections and interactions, the convergence and divergence of groups, historical associations and developing trends.

In this paper, I will introduce Sociolinguistic research and the nature of sociolinguistic field techniques and sample design. I will argue that socially embedded data is critical for analyzing and discovering social meaning. Then, I will summarize the findings of several case studies. What does the use of a 3rd singular morpheme -s, as in (1), tell us about the history and culture of a community (Tagliamonte 2012, 2013)? How is quotative be like, (2), spreading in geographic space (Tagliamonte to appear)? What is the mechanism that underlies linguistic change (Tagliamonte & D'Arcy 2009) and by extension cultural trends and projections?

1. The English people speaks with grammar.
2. I was *like*, "Hey how are you going?" And hes *like*, "Im fine."

Using sociolinguistic datasets, the answers to these questions have successfully addressed prevailing puzzles and offered solutions to real world problems. However this type of research is only be as good as the quality of the data, the capability of the technologies for extracting and analyzing what is important, and the relevance of the socially cogent and statistically sound interpretations. I will argue that Sociolinguists and Computational Scientists could be powerful allies in uncovering the complex structure of language data and in so doing, offer unsurpassed insight into varying human states and conditions.

References

William Labov. 1994. *Principles of Linguistic Change: Volume 1: Internal Factors*. Blackwell.

William Labov. 2001. *Principles of Linguistic Change: Volume 2: Social Factors*. Blackwell.

Sali A. Tagliamonte and Alexandra D'Arcy. 2009. Peaks beyond phonology: Adolescence, incrementation, and language change. *Language*, 85(1):58–108.

Sali A. Tagliamonte. 2012. *Variationist Sociolinguistics: Change, Observation, Interpretation*. Wiley-Blackwell.

Sali A. Tagliamonte. 2013. *Roots of English: Exploring the History of Dialects*. Cambridge University Press.

Sali A. Tagliamonte. To appear. System and society in the evolution of change: The view from Canada. In E. Green and C. Meyer, editors, *Faces of English*. De Gruyter-Mouton.