Closer Still to a Robust, All Digital, Empirical, Reproducible Sociolinguistic Methodology

Christopher Cieri & Stephanie Strassel (*University of Pennsylvania, Linguistic Data Consortium*)

Advances in human language technologies and interfaces for coding sociolinguistic variables have brought some research groups to the point of a robust, digital, reproducible sociolinguistic methodology.

Our vision for this methodology includes not only digital data collection but also time aligned transcription as an index to speech that can be searched for variables of interest optionally aided by pronouncing dictionaries or letter to sound rules. Coding decisions are still made by humans though the potential for partial automation exists. Variables, and coding practice are described fully to permit replication by others on the same or comparable data. Decisions are tracked in databases so that individual data points, dots on a scatter plot or examples in a paper, can be tracked backed to the original recordings.
Ideally the data is also publicly accessible.

A number of recent studies have instantiated significant parts of this vision for example, Minnick-Fox’s dissertation on Spanish s-lenitition. More recently the Phanotics project, undertaking sociolinguistic coding to support forensic speaker recognition, has used publicly accessible data, created transcripts and used speech recognition technology to align transcripts to audio at the word and phone levels. Phanotics has also carefully defined specifications to support coding at multiple independent sites and preserved the links that match data points to offsets in the audio recordings. After reporting on Phanotics activities, including results to date, this paper will discuss available tools that bring the same capabilities to other research groups and changes in practice that would maximize the benefit of these tools.

Reference

Minnick-Fox, Michelle, 2006, Usage-based effects in Latin American Spanish syllablefinal /s/ lenition, University of Pennsylvania, Doctoral Dissertation.