Mark Liberman

The problems of scale in language documentation

The project of documenting endangered languages must come to terms with the problems posed by the amount of primary data that is needed. For example, calculations of several different kinds converge to estimate the need for text at about 10 million words per language. In spoken form, this will be more than 50,000 hours of recordings, a scale that is two or three orders of magnitude greater than current practices are yielding. And the problem of recording is dwarfed by the problems of transcription and analysis.

Is it possible for language documentation efforts to meet this challenge? We need effective applied research of several kinds: to determine how much primary data is really needed for adequate documentation; to radically increase the amount of primary data that is collected and analyzed for a given amount of investment; and perhaps to increase the documentary value of collected data, so as to decrease the amount of primary data that is needed to reach an adequate level of documentation.

This is a very difficult set of problems, but unless they are solved, the goal of language documentation will remain out of reach. Any solution will require significant improvements in project design, organization and productivity, with close collaboration among computational linguists, field linguists and the affected speech communities. The scope of the problem also motivates more extensive involvement of more of the world's linguists, in a process that would bring benefits to all parties.

In this talk, I'll focus on some of the challenges and opportunities for computational linguistics, broadly construed.