Mitch Marcus
University of Pennsylvania

Title: Code switching: A new tool for corpus collection and its application

Abstract:

Codeswitching, the concurrent use of more than one language by a single speaker, is of linguistic interest both for its sociolinguistic implications and for what formal constraints on codeswitching reveal about syntax. Because codeswitching occurs largely within informal genres, corpora for the study of this phenomenon derive from transcriptions of sociolinguistic interviews or from transcription of small sets of examples and are of relatively limited size. Just as the availability of large annotated corpora have facilitated research in historical linguistics, large annotated corpora of codeswitching across a range of languages would be highly useful.

In joint work with Constantine Lignos, we have developed a tool that labels codeswitched and monolingual Twitter messages (tweets) and labels words within those tweets for source language with reasonable accuracy. The design of the algorithm behind this tool combines statistical methods and linguistic insights in novel ways to improve the accuracy of the tool. I will discuss both the simple algorithms behind this tool and a very simple method that allowed a reasonably large high precision, low recall test collection for initial algorithm development to be collected with only a few hours of work. Finally, I will present work by Lignos that uses this corpus to empirically evaluate the accuracy of proposed formal constraints on codeswitching.