Richard Littauer

Constructing Corpora for Low Resource Languages from Social Media

Minority and endangered languages often have little or no extant corpora usable by the linguistics or computational linguistics community. Compiling corpora is the first step towards providing tools and ways for users to speak their language outside of their spoken communities, such as on the internet. In this talk I will talk about the possibility of using social media, such as Twitter and Facebook, for corpus creation. I will discuss legal issues, present work I have done on the language Rangi, and showcase an XML schema appropriate for storing corpora from social media websites.