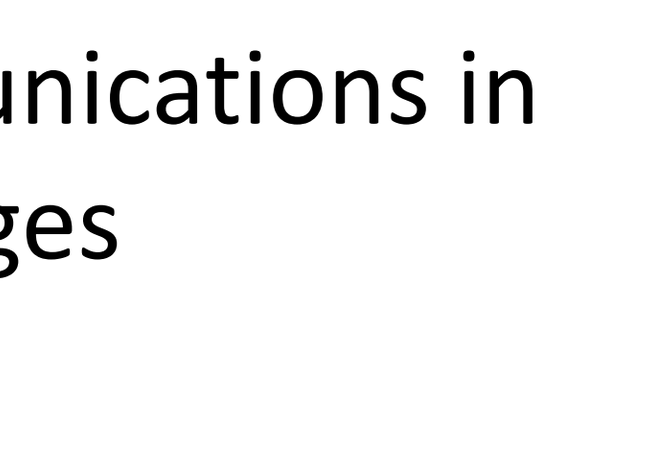
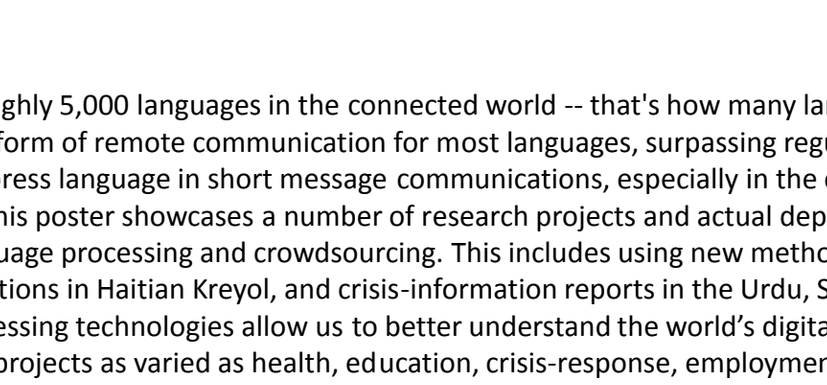
Understanding Communications in 5,000 languages

Robert Munro

The recent global proliferation of technology, cell phones in particular, means that there are roughly 5,000 languages in the connected world -- that's how many languages you could find at the other end of your phone right now. Text-messaging (SMS) is the most popular form of remote communication for most languages, surpassing regular mail, email, and actual phone calls. However, very little is known about the nature of how people express language in short message communications, especially in the context of non-standardized spellings, varying literacy, and frequent code-switching between languages. This poster showcases a number of research projects and actual deployments that seek to triage communications in less-resourced languages, leveraging advances in natural language processing and crowdsourcing. This includes using new methods to process health-related messages in the Chichewa language of Malawi, emergency response communications in Haitian Kreyol, and crisis-information reports in the Urdu, Sindhi and Pashto languages of Pakistan. For all three contexts, it is shown that new natural language processing technologies allow us to better understand the world's digital linguistic diversity and in turn how we can use the same technologies to aid the speaker communities in projects as varied as health, education, crisis-response, employment, and access to market information.

Dalila: I need Thomasssin Apo, please
Apo: Wait
Apo: Kenscoff Route: Lat: 18.4957, Long:-72.3184
Apo: The area after Petion-Ville and Pelerin 5 is not on Google Map.
Apo: We have no streets name
Dalila: Apo, thanks for your help
Apo: you are welcome. I know this place like my pocket :)
Dalila: thank God u was here

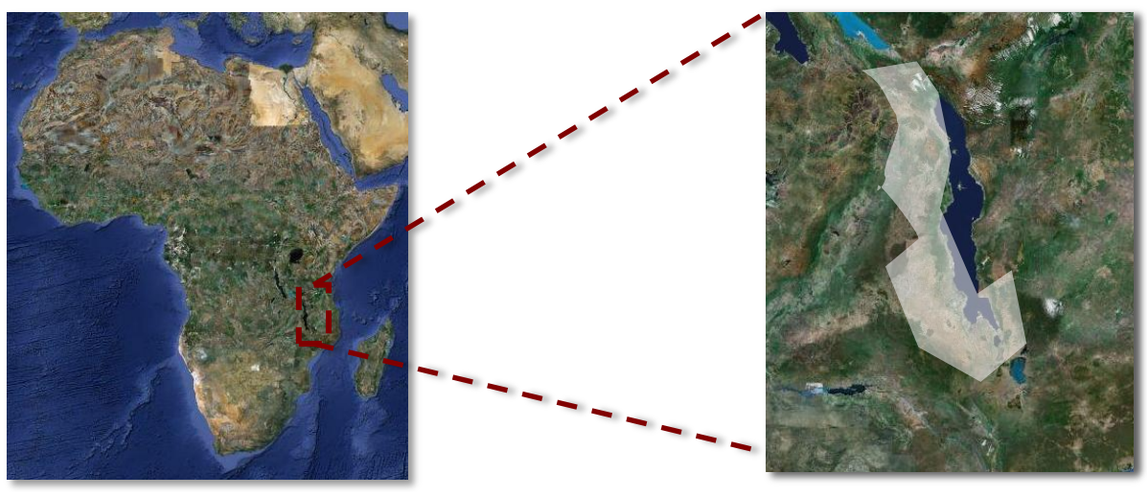


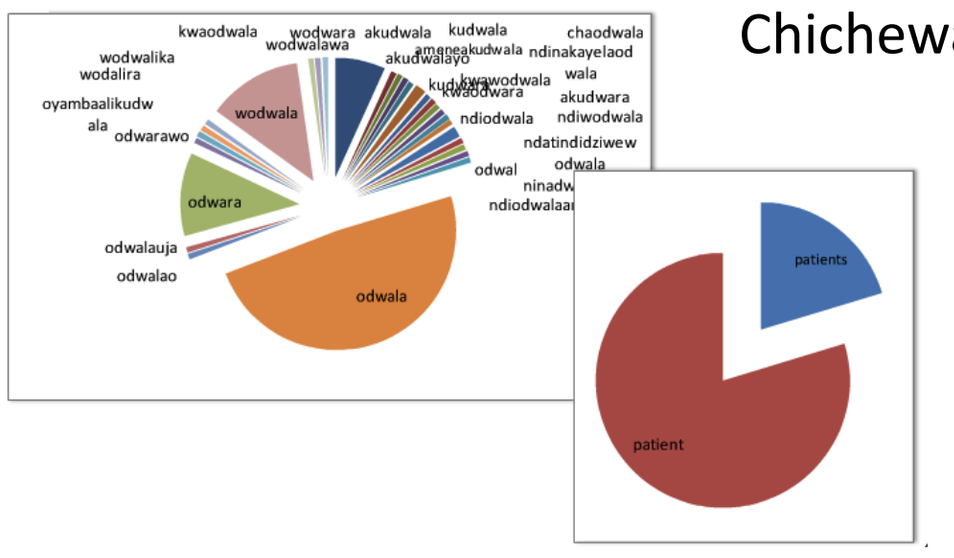


Haitian Kreyol

Urdu, Pashto, Sindhi

Karachi: Project Madad: Need Volunteers for data entry for relief inventory
'Karachi: Project Madad: rahat soochi keliye atha pravishti karne keliyae svayanasevak ka zaroorat hai.'





Chichewa

ndimmafuna manthwala
('I currently need medicine')
↓
ndimafuna mantwala
↓
ndi-ma-fun-a man-twala
↓
ndi-ma-**fun**-a **man-twala**
↓
ndi -fun man-twala
("I need medicine")
Category = "Request for aid"

ndi kufuni mantwara
('my want of medicine')
↓
ndi kufuni mantwala ← 1) Normalize spellings
↓
ndi-ku-fun-i man-twala ← 2) Segment
↓
ndi-ku-**fun**-i **man-twala** ← 3) Identify predictors
↓
ndi -fun man-twala
("I need medicine")
Category = "Request for aid"

1 in 5 classification errors with raw messages

1 in *20* classification error post-processing. *Improves* with scale.