Towards Language Documentation 2.0

Imagining a Crowdsourcing Revolution

Mat Bettinson - the University of Melbourne
There is presently not the level of activity that we’d ideally like to see - language loss despite our best intentions.

Against this, we are presently observing an explosion of self-documentation in the social web.

In this talk I present a thought experiment of a fictional crowdsourcing utopia.

The goal is to stimulate debate on how we may move closer to language documentation 2.0.
New technology and methods
In a few short years...

- **Basic Oral Language Documentation**
- **BOLD PNG**
- **Aikuma**

Reiman (2010)

Bird (2010)

Bird, Hanke, Adams & Lee (2014)
You may say I’m a dreamer but I’m not the only one...

“And now we can start to dream of seeing collaborative teams that start to gather in the village, process in the local university and share quality material with international linguists and archives”

John Hatton - ICLDC3, 2013

SayMore - Language Documentation Productivity
The thought experiment
In fictional Bluegreen land...

... In the not too-distant future, in the town of blue.
The thought experiment
In fictional Bluegreen land in the future...

- Ambreen gets a message on her phone

Come to the right side of the powerpoint slide, old men are telling stories!
We arrive in the village... is that a phone?

Rural Cambodia
Julia Maudlin (Flickr, creative commons license)
Papua New Guinea

Aiyura Valley, Eastern Highlands of Papua New Guinea
by Kahunapule Michael Johnson (Flickr, creative commons license)
Mobile-cellular subscriptions per 100 inhabitants

10 most language-diverse countries in the world
ITU: ICT data and statistics explorer
Thought experiment cont.
The story of the mountain lion

- Ambreen’s friend Saniu is recording a video of the elders telling a story about the mountain lion
- Ambreen and Saniu don’t understand what’s being said (but they like the lion sounds)
- If they record them now, later on they will be able to...
Thought experiment cont.
ROAR! And then he ate...

- ... watch the recording with sound and words in the new language!
App-based fieldwork
Somali urban fieldwork example

- Raw recording
- Respeaking
- Translation
Thought experiment cont.
Meanwhile in a local university

- Abnam coordinates a language documentation project
- He asked Saniu (via the network) to record the mountain lion story: *network fieldwork*
- The material needs to be processed but we have long since moved past one person doing such time consuming work...
Network fieldwork
Another example from Somali

- The smartphone app Aikuma gained network ‘backup’
- Participant has a mobile with a SIM card & data plan
- I was then able to review the work (which was a huge help!)

- Original respeaking recording:

- “Slower in a quiet place please”: 
Thought experiment cont.
Crowd-curation

- Sumkhuu is interested in his ancestral language and he’s not bad with computers
- He is one of several people contributing towards the processing and curation of materials coordinated by Abnam
- He’s not a linguist but he can do written orthographic transcriptions
- His work is made easier by utilising computational techniques, such as automatic transcription, lexicon-building, ‘forced alignment etc’
Thought experiment cont.
Finally, somewhere else in Bluegreen land

- Bimo is some random guy in Bluegreen land
- While browsing Facebook and saw his friend post a video of a mountain lion story
- Bimo speaks Kita pretty well, so he can tell the story isn’t translated quite right
- He clicks on the Facebook post and ends up on the language network
- He is now aware of the Kita revitalisation project and decides to take part
Web 2.0
A new person on the web

- For every new village and town that joins the internet, it’s the Web 2.0 that they will first experience
- There are some sparse accounts of communities adopting these tools for language activism and revitalisation (Campbell & Huck, 2013)
- Looking for further evidence...
Case study in Taiwan
An ‘endangered’ Formosan language

► Facebook is automatically making ‘language’ pages taken from Wikipedia
► Saisiyat (language code ‘xsz’) is spoken by an ethnic group with a population of less than 5,000
► Sure enough, there’s a Facebook page:

► 260 Facebook users have said they speak the language!
Case study in Taiwan
An ‘endangered’ Formosan language

- I can view the speakers, look at their profiles and find Saisiyat Facebook groups
Phil the grad student has a meeting with his potential supervisor Norman.

Norman asked: “Why do yet another PhD on a European language when you can work on nasal discourse evidentials in Kitapuripurinamca?”

Phil: “Fieldwork? I’ve heard there are mosquitos!”

Norm: “Fear not, fearful Phil, I’ll hook you up with Abnam in Bluegreen land.”

Phil writes his dissertation on Kita (with no mosquitoes) while contributing towards the documentary project at large.
How do we get there?
Recent work

- Dunham, Cook & Horner (2014) considered software requirements for ‘collaborative fieldwork software’
- Birch et al. (2013) raised a number of challenges for ‘app-based’ crowdsourcing
- Major commonalities:
  1. The importance of engaging end-users: User-friendly UIs, rewards, game-ification and so on
  2. The importance of data curation, either ‘collaboratively’ or via crowd-curation
  3. The need for non-trivial permissions balanced against the somewhat conflicting goal of encouraging sharing and collaboration
- Lots of helpful experience from current tools such as SayMore (particularly with attention to usability)
You might have missed this
A hint of the dream in Taiwan

- Klokah.tw (indigenous language paradise) is an inspiring glimpse of the future:
Conclusion

- *Language Documentation 2.0* is broadly conceived as a Web 2.0-like approach to collective documentary activity

- The social web provides a spectacular opportunity to engage and recruit

- The fieldwork landscape is changing rapidly: ever more locations are increasingly under the digital footprint

- We should not be surprised that communities have digital experience and *expectations*

- We hypothesize that network fieldwork offers an opportunity to elevate endangered languages from a ‘marginal position in linguistics’ (Newman, 1998, 2003)

- Some previous discussion of software requirements but we need more research (and practice) on the *network* of multiple software tools
Thanks for listening!

Mat Bettinson
mbettinson@unimelb.edu.au
@sinomat

For detail on crowdsourcing language with mobile devices, please see Steven Bird’s talk 9am tomorrow
http://www.lp20.org
References


