Towards Language Documentation 2.0

Imagining a Crowdsourcing Revolution

Mat Bettinson – the University of Melbourne

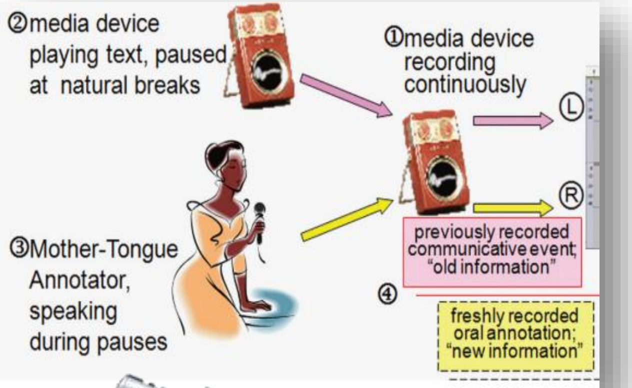
Introductory Context
Language loss

- 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

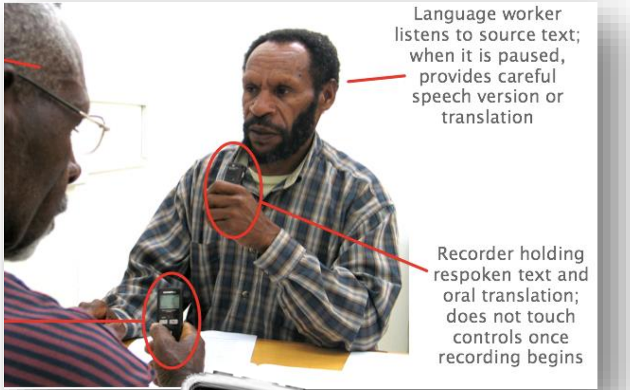
In a few short years…

- Basic Oral Language Documentation


②media device playing text, paused at natural breaks
①media device recording continuously
Ⓛ
Ⓡ
③Mother-Tongue Annotator, speaking during pauses
previously recorded communicative event; "old information"
④
freshly recorded oral annotation; "new information"

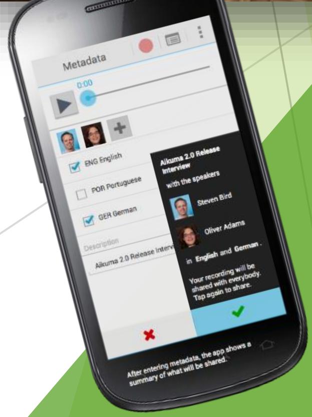
Reiman (2010)

- BOLD PNG


Language worker listens to source text; when it is paused, provides careful speech version or translation
Recorder holding respoken text and oral translation; does not touch controls once recording begins

Bird (2010)

- Aikuma



Aikuma 2.0 Release Interview with the speakers Steven Bird Oliver Adams in English and German.

Your recording will be shared with everybody. Tap again to share.

After entering metadata, the app shows a summary of what will be shared.

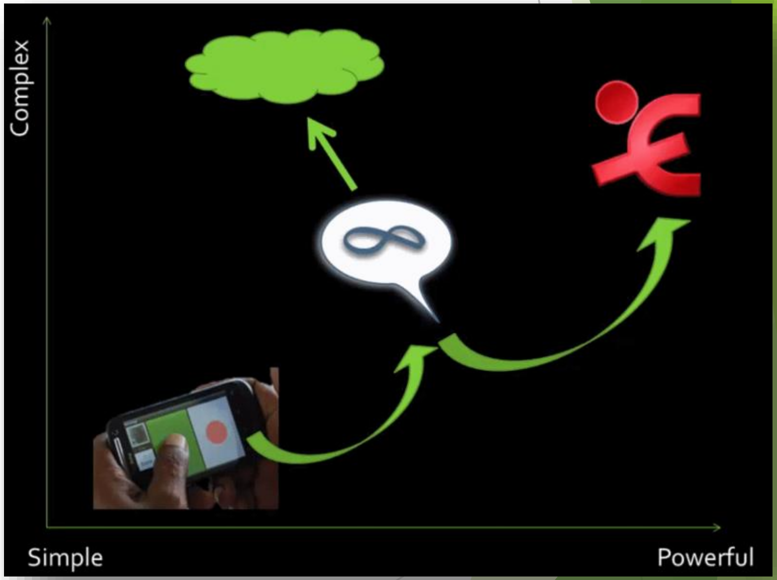
Bird, Hanke, Adams & Lee (2014)

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



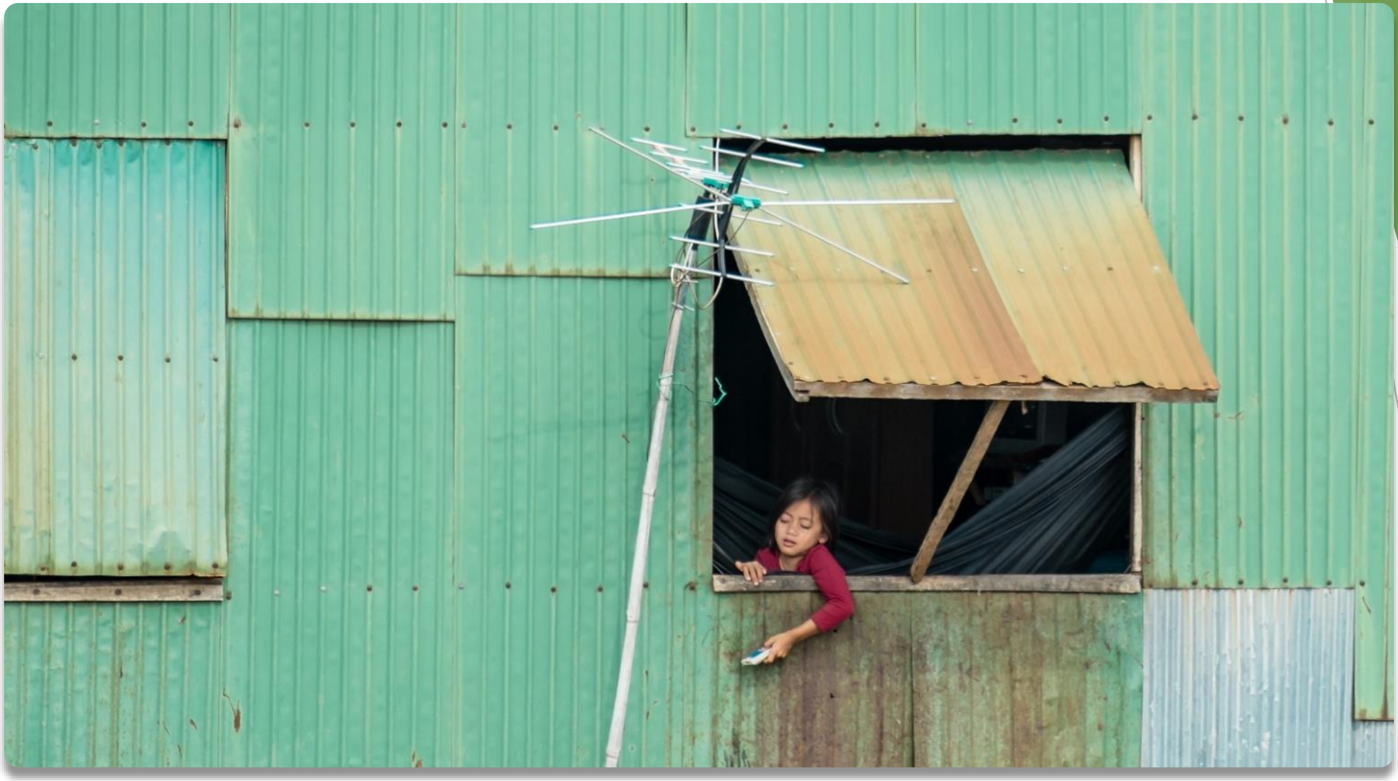
The thought experiment
In fictional Bluegreen land…

... 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?

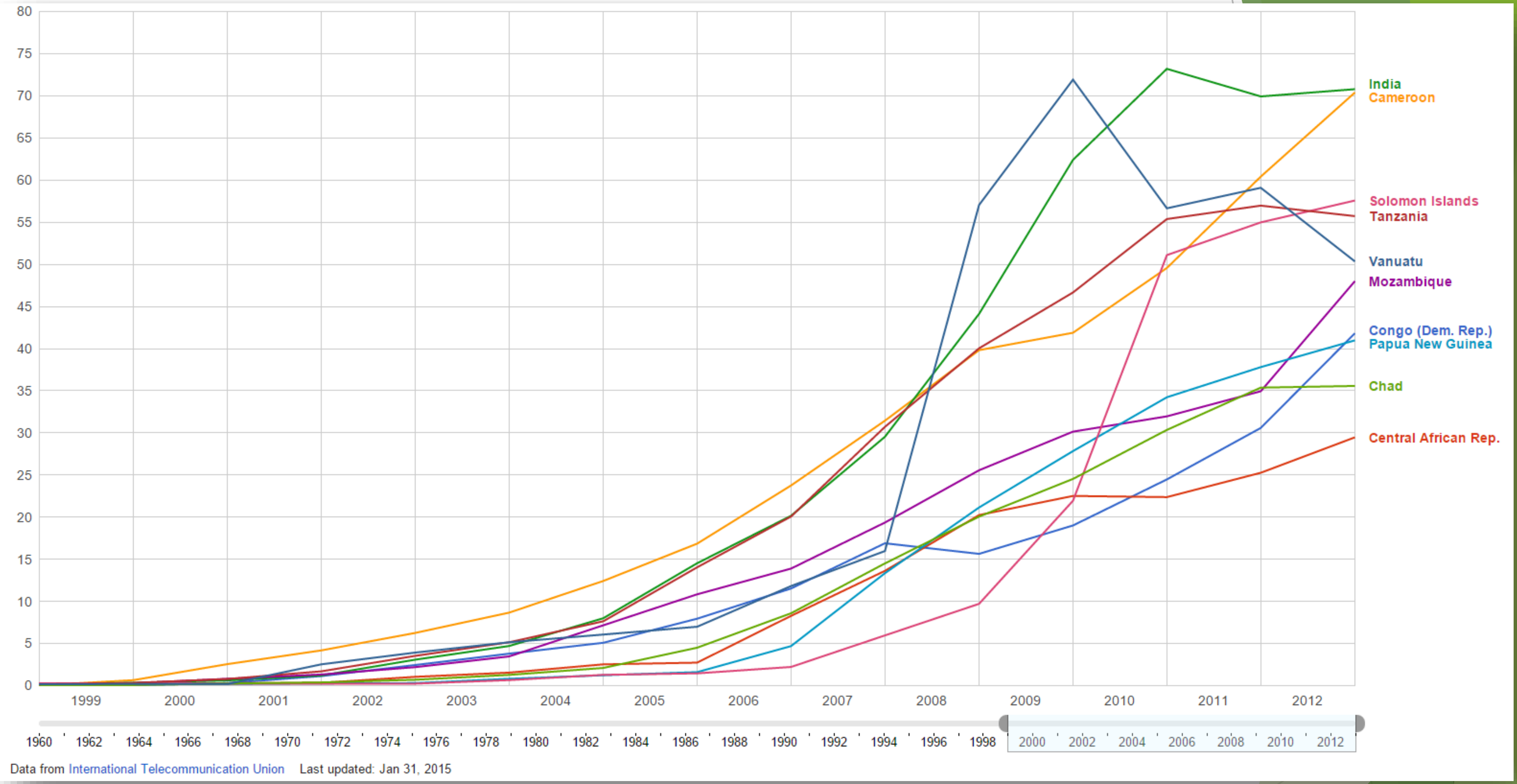
Rural Cambodia
Julia Maudlin (Flickr, creative commons license)

Digicel
Top Up Here!
Digicel
Charge your mobile phone here!
Charge your mobile phone here!
Digicel

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



Data from International Telecommunication Union   Last updated: Jan 31, 2015

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!

AND THEN HE ATE
!!!

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…

SANVI, I NEED MORE LION STORIES!
ALSO, PLEASE ASK THEM TO ROAR SLOWLY AND CLEARLY.

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

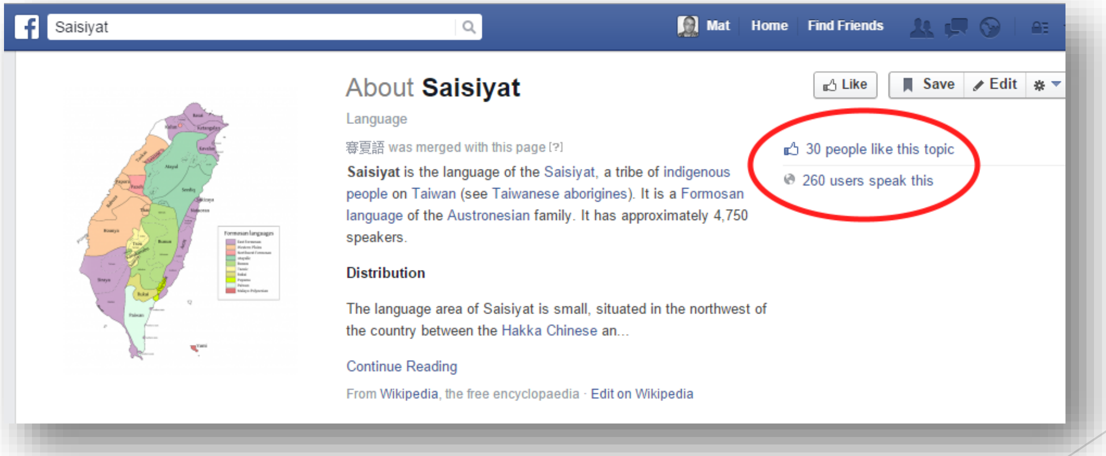
Web 2.0

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:



About **Saisiyat**

Language

賽夏語 was merged with this page [?]

Saisiyat is the language of the Saisiyat, a tribe of indigenous people on Taiwan (see Taiwanese aborigines). It is a Formosan language of the Austronesian family. It has approximately 4,750 speakers.

Distribution

The language area of Saisiyat is small, situated in the northwest of the country between the Hakka Chinese an...

Continue Reading

From Wikipedia, the free encyclopaedia · Edit on Wikipedia

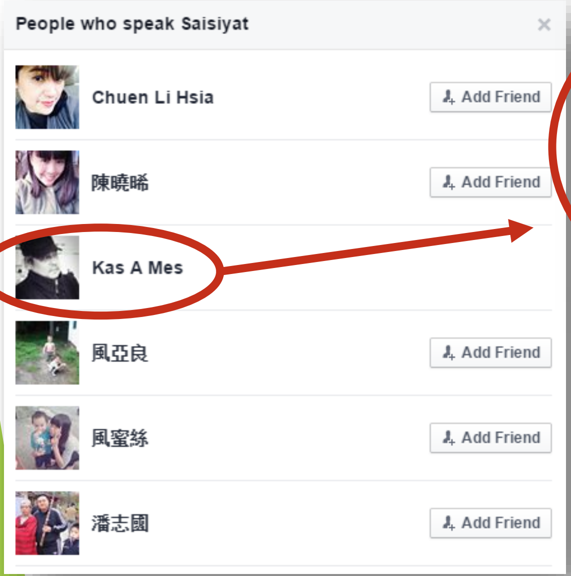
30 people like this topic

260 users speak this

- 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



People who speak Saisiyat ×

Chuen Li Hsia — Add Friend

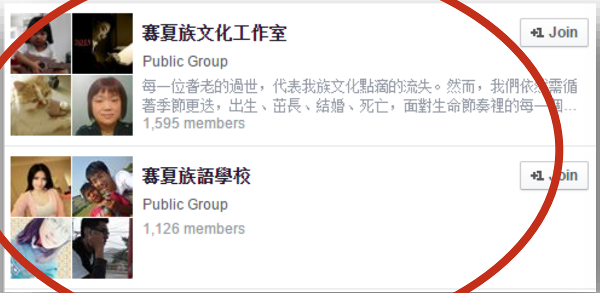
陳曉晞 — Add Friend

Kas A Mes

風亞良 — Add Friend

風蜜絲 — Add Friend

潘志國 — Add Friend



賽夏族文化工作室 — +1 Join

Public Group

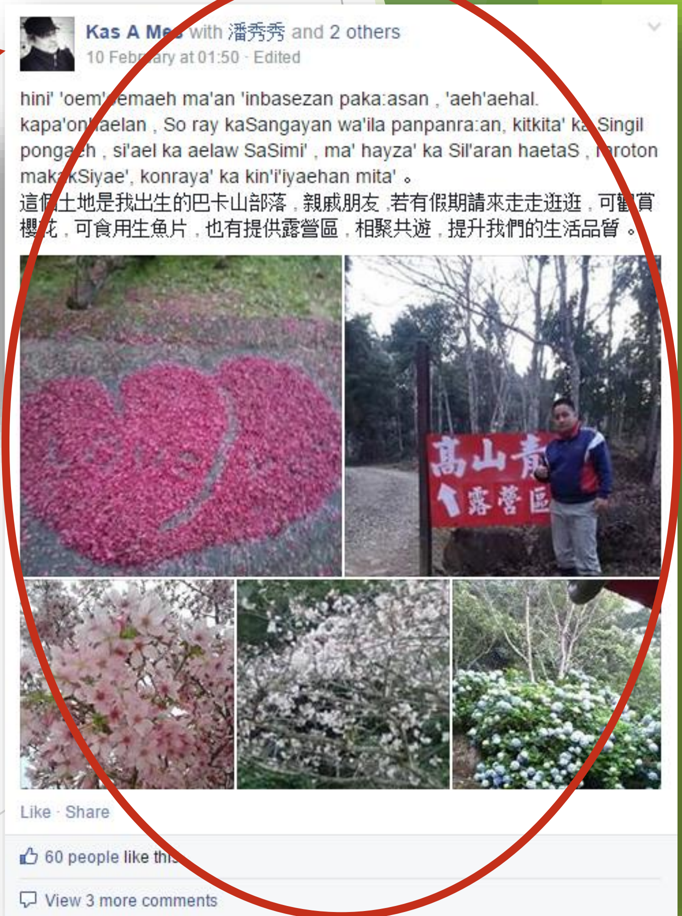
每一位耆老的過世，代表我族文化點滴的流失。然而，我們依然需循著季節更迭，出生、茁長、結婚、死亡，面對生命節奏裡的每一個…

1,595 members

賽夏族語學校 — +1 Join

Public Group

1,126 members



Kas A Mes with 潘秀秀 and 2 others

10 February at 01:50 · Edited

hini' 'oem'oemaeh ma'an 'inbasezan paka:asan , 'aeh'aehal. kapa'onhaelan , So ray kaSangayan wa'ila panpanra:an, kitkita' ka Singil pongaeh , si'ael ka aelaw SaSimi' , ma' hayza' ka Sil'aran haetaS , naroton makakSiyae', konraya' ka kin'i'iyaehan mita' 。

這個土地是我出生的巴卡山部落，親戚朋友，若有假期請來走走逛逛，可觀賞櫻花，可食用生魚片，也有提供露營區，相聚共遊，提升我們的生活品質。

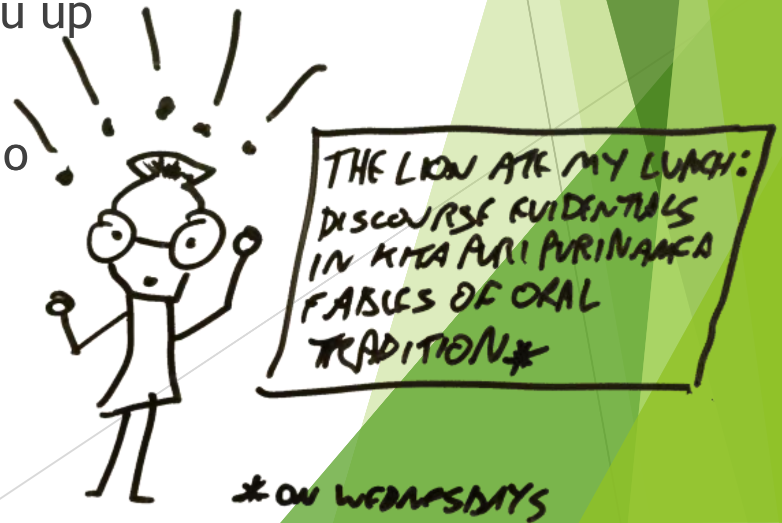
Like · Share

60 people like this

View 3 more comments

Thought experiment cont.
In a foreign university near you

- 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


THE LION ATE MY LUNCH: DISCOURSE EVIDENTIALS IN KITA PURI PURINAMCA FABLES OF ORAL TRADITION*
*ON WEDNESDAYS

How do we get there?
Recent work

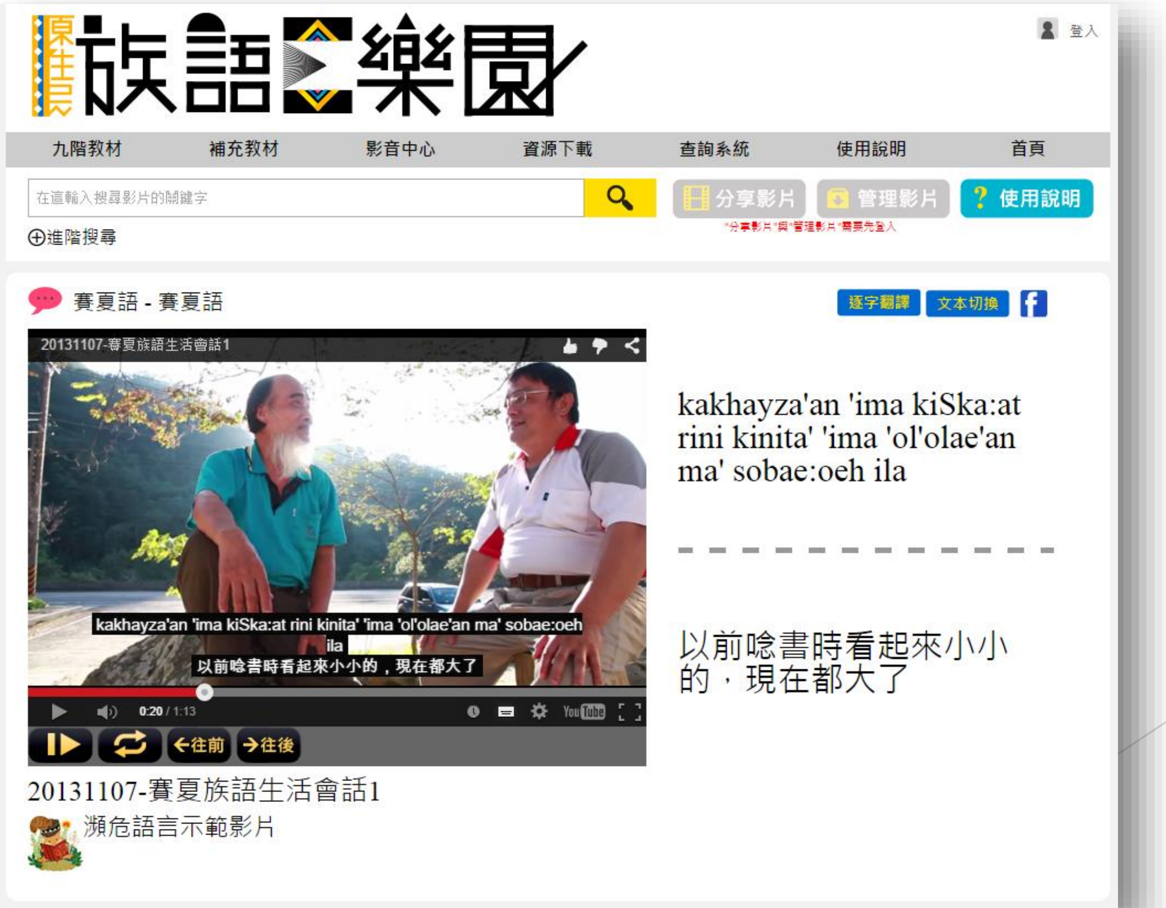
- Dunham, Cook & Horner (2014) considered software requirements for ‘collaborative fieldwork software’
- Birch et al. (2013) raised a number of challenges for ‘app-based’ crowdsourcing
- Major commonalities:

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:



賽夏語 - 賽夏語

20131107-賽夏族語生活會話1

kakhayza'an 'ima kiSka:at rini kinita' 'ima 'ol'olae'an ma' sobae:oeh ila

以前唸書時看起來小小的，現在都大了

kakhayza'an 'ima kiSka:at rini kinita' 'ima 'ol'olae'an ma' sobae:oeh ila

以前唸書時看起來小小的，現在都大了

20131107-賽夏族語生活會話1

瀕危語言示範影片

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

Birch, B., Drude, S., Broeder, D., Withers, P., & Wittenburg, P. (2013). Crowdsourcing and apps in the field of linguistics: Potentials and challenges of the coming technology.

Bird, S., Hanke, F., Adams, O. & Lee, H. (2014). Aikuma: A Mobile App for Collaborative Language Documentation. Workshop on the Use of Computational Methods in the Study of Endangered Languages, Baltimore, USA

Campbell, B & Huck, J. (2013). Social Media as a Tool for Linguistic Maintenance and Preservation among the Mapuche. Proceedings of the 2013 LAGO Graduate Student Conference "Decolonizing the Americas", Tulane University.

Dunham, J., Cook, G., & Horner, J. (2014). LingSync & the Online Linguistic Database: New models for the collection and management of data for language communities, linguists and language learners. In Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages (pp. 24-33).

Cathcart, M., Cook, G., Deering, Manyakina, Y., McCullock, G., Noguchi, H. (2012). LingSync: A free tool for creating and maintaining a shared database for communities, linguists and language learners. In Robert Henderson and Pablo Pablo, editors, Proceedings of FAMLi II: workshop on Corpus Approaches to Mayan Linguistics 2012, pages 247–250.

Hatton, J. (2013). SayMore: Language documentation productivity. Third International Conference on Language Documentation and Conservation (ICLDC). http://hdl.handle.net/10125/26153

Newman, P. (2003). The endangered languages issue as a hopeless cause. In M. Janse & S. Tol (Eds.), Language death and language maintenance: Theoretical, practical and descriptive approaches Vol 240 (pp. 1–14). John Benjamins Publishing.

Reiman, D. W. (2010). Basic oral language documentation. Language Documentation & Conservation, 4, 254–268.