Language archives and new research methods in support of small languages

Nick Thieberger
PARADISEC, School of Languages and Linguistics, University of Melbourne, Australia

Archiving is a critical part of language documentation, building citable forms of primary records on which anlyses are built, but access to heritage language records is only possible if they are described in standard metadata terms, are digital, and can be found on the web.

Existing analog recordings need to be located and digitised before the tapes become unplayable or the playback machines are no longer available. This has been a major focus for PARADISEC, with material from over 830 languages.

Records made by linguists or musicologists are born digital, but, even then, there is a need for training to create records that are of good quality. The researcher typically knows more contextual information than they have written in a catalog entry, so we have to build better methods to encourage them to build better descriptions of their collections. Further, even with training, there need to be incentives within academia that reward the production of well-described collections of primary records. In this way we can look forward to better descriptions of collections held in archives.

Our catalog exports an XML feed in standard protocols for internet harvesters via the Open Language Archives Community. Archives that subscribe to OLAC then also feed a global resource that shows what records are available per language, a very useful index of existing documentation.

Looking to the future we can envisage cheaper storage and faster access via a range of mobile devices with automated identification of segments in media files and alignment of transcripts with their media source. Mobile devices will also create new recordings. Online access to records lends itself to crowdsourcing for annotations if the records are presented in a suitable format. However, without an effort to make it easier to archive such records there is a potential for them to be lost.

Sites

Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC): http://paradisec.org.au
Language Documentation & Conservation Journal: http://www.nflrc.hawaii.edu/ldc/ /
DELAMAN, project Lost and Found: http://www.delaman.org/project-lost-found/

References

Bird, Steven and Gary Simons 2003. “Seven dimensions of portability for language documentation and description”. *Language* 79:557-582. (http://www.sil.org/~simonsg/preprint/Seven%20dimensions.pdf)

Chang, Debbie 2010. Taps: Checklist for Responsible Archiving of Digital Language Resources. MA thesis, Graduate Institute of Applied Linguistics

Thieberger, Nicholas 2004. Documentation in practice: Developing a linked media corpus of South Efate. Peter Austin (ed*). Language documentation and description*, Volume 2. London: Hans Rausing Endangered Languages Project, SOAS. 169-178. (http://repository.unimelb.edu.au/10187/2199)

In addition to the abstract, please include the following information:
NAME(S): Nick Thieberger
TITLE: **Language archives and new research methods in support of small languages**
INSTITUTION: University of Melbourne, Australia
E-MAIL: thien@unimelb.edu.au
ADDRESS: PARADISEC / School of Languages and Linguistics, University of Melbourne, Parkville, Vic 3010, Australia

TEL: +61 3 8344 8952