18

Language Documentation & Conservation Special Publication No. 3 (August 2012):
Potentials of Language Documentation: Methods, Analyses, and Utilization,
ed. by Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann,
Dagmar Jung, Anna Margetts, and Paul Trilsbeek pp. 129–134
http://nflrc.hawaii.edu/ldc/sp03/
http://hdl.handle.net/10125/4527

Using language documentation data in a broader context

Nick Thieberger
University of Melbourne / PARADISEC

On the one hand we have never seen as much fieldwork and recording of small and endangered languages as we have over the past decade. On the other hand linguists are now also much more aware of the need to create records that can be reused by the people we record and that will still be available for their descendants. Our own descendants, the future researchers who will use our records, will also need to be able to find and make use of our research. The fragility of digital records means we need to pay attention to their curation over time and create suitable repositories if they do not already exist. In order for these aims to be achieved, we need to establish work practices now that allow the data to move easily from creation to the archive and to community use.

1. How can language documentation data be utilized in a broader context? There are obviously many ways in which language documentation data can be used, but first they must be created in ways that permit ready reuse. That is, the data must exist in formats that allow them either to be used immediately or to be converted to a usable form without too much effort. Reworking data by hand is too time consuming, and '[t]he amount of data produced far exceeds the capabilities of manual techniques for data management.' (Borgman 2007: 6). To automate handling of linguistic data, it needs to be created in appropriate structures to begin with. This implies that we have tools that produce output in reusable forms and that linguists have training in what it means to create reusable data. We need to participate in broader initiatives in the Humanities that are leading to the development of the necessary infrastructure to house our research outputs in the longterm. Beagrie et al. (2008: 22) suggest that community-wide agreement on standards enhances the preservability of the data generated by that community. In linguistics we are fortunate to have some established standards (e.g., Leipzig glossing rules[1], OLAC[2], and IMDI[3], among others), but there is variable uptake of these standards and a corresponding lack of understanding of the reasons for using them. This paper addresses ways of using linguistic data based on an assumption that more use can be made of data that is created adhering to such standards.

[1] http://www.eva.mpg.de/lingua/pdf/LGR08_09_12.pdf

[2] http://www.language-archives.org

[3] http://www.mpi.nl/IMDI/

The creation of this data requires specialized knowledge of data formats, which should be provided by training researchers in documentation methods. Alternatively, linguists may be able to use tools that provide well-formed data and so bypass a need to really understand the underlying form of the data they are creating. This is typically the way in which most current linguistic data is created, using tools such as Elan[4] or Toolbox[5] and their output text files rather than proprietary formats as produced, for example, by Microsoft software. With training in the use of the tools, linguists can generate the kinds of material they need and at the same time allow the data to be archived and reused. Such training has been provided at organized summer schools or training workshops or intensive workshops associated with linguistic conferences (such as those run by InField[6], RNLD[7], LLL[8], ELDP[9], or DoBeS[10]).

While those working in projects funded by ELDP or DoBeS routinely receive training, the majority of language documentation is still being carried out by linguists trained in descriptive fieldwork and not in documentation methods. For example, Newman (1992, 2005) reports 34 US departments running field methods courses, most of which it can be assumed did not include documentary methods such as time-aligned transcriptions of the primary recordings, metadata catalogs of files generated in the project, annotation of files with interlinear glossed text and lexicons in open and reusable formats, or preparation of the material for long-term archiving.

However, if we take into account that there were some 230 papers at the 2nd International Conference on Language Documentation and Conservation and 180 papers at the LLL conference in 2009 there would appear to be quite a great deal of activity in language documentation projects. We can assume from this that there are at least 100 current fieldwork-based linguistic projects so, if we extrapolate back with a conservative estimate of 50 new projects per year since 1960, there should be some records of 2,500 small languages. Without a concerted effort to describe and curate the outputs of this research, many of the primary recordings prepared by these linguists have been or will be lost, if they ever existed to begin with (keeping in mind that, until recently, it was quite acceptable – and even advocated by Dixon (2006) – to use as little technology as possible, to make few or no field recordings and to make no provision for their long-term accessibility).

Some linguists have voiced concerns about the use of new technology in recording language material. The concerns range from the alleged neo-colonial nature of the use of high technology tools in third-world settings (Aikhenvald 2007) to the notion of commodification of language in an 'audit culture' (Dobrin et al. 2009) and include a critique of what they perceive as an excessive focus on the tools rather than on the language. It is tempting to ignore these arguments in the hope that they will be overtaken by the practicalities of doing fieldwork, and in the fear that reproducing them will give them more credence than they should receive. Unfortunately, it is still necessary to point out that any professional needs to

---

4 http://www.lat-mpi.eu/tools/elan/manual/

5 http://www.sil.org/computing/toolbox/

6 http://darkwing.uoregon.edu/~spike/Site/InField_2010.html

7 http://www.rnld.org

8 http://www.ddl.ish-lyon.cnrs.fr/AALLED/Univ_ete/Summer_school.html

9 http://www.hrelp.org/events/workshops/

10 http://www.mpi.nl/DOBES/training_courses/

use current technologies for their work, and that for linguists this means learning and using new tools and methods for fieldwork and analysis.

2. How must these digital data be stored, represented, and made accessible by the archives? Accessibility is based on locatability of the material in the collection. This relies on catalogs that use standard terms and are accessible via normal search mechanisms (e.g. Google, or the Open Archives Initiative (OAI)[11]). Non-compliant repositories – those whose catalogs do not conform to the normal standards (e.g., OLAC) of our research community – should be encouraged to conform. The benefit of conforming is that federated searches will include these repositories and we will thus begin, as a community of researchers, to create a dynamic documentation index of archived information about the world's languages, as provided by OLAC, and then harvested by any similar project (e.g. the Clarin Virtual Language Observatory[12] or ELCat[13]).

At the time of writing, two large-scale funding programs, ELDP[14] and DoBeS[15], have facilitated a great deal of research. The ELDP has funded 216 projects and DoBeS has funded 51. However, the ELDP archive, ELAR, contains (at the time of writing) just 110 collections, which raises the question, if a funding body like the ELDP cannot get all of its grantees to deposit in an archive in a timely fashion (or at all), how can unfunded researchers be expected to make use of an archive? One solution could be better training and raising the awareness of researchers of the need to archive, including appealing to their responsibility to make records for others to access in future. Another approach would be to make it as easy as possible to deposit in archives by use of new metadata creation tools (like, for example, ExSite9[16] or Arbil[17]).

For objects in repositories that will not be able to conform (that is, state libraries, archives, or similar institutions), a service could be built that indexes language material in these collections and assigns simple descriptors (like standard language codes) with links to the URI of the object. For example we can create a record in the PARADISEC catalog linking to an item found on the Anglican Church website that is written in Raga, a language from Vanuatu[18]. PARADISEC approached the website owners to ask if they would consider archiving their primary material, but had no response from them. Trusting that the pages on the Anglican Church website have some persistence, the links from the PARADISEC catalog will allow this item to be found in a federated search of all OAI archives.

The OLAC search page[19] provides a targeted search tool for language material, but is only as good as the collections it harvests. OLAC currently includes forty digital language

[11] www.openarchives.org

[12] http://catalog.clarin.eu/ds/vlo

[13] http://www.endangeredlanguages.com/

[14] Endangered Languages Development Programme, http://www.hrelp.org/languages/

[15] http://www.mpi.nl/DOBES

[16] ExSite9 (formerly FieldHelper) is a tool being produced in Sydney for creating standard metadata via drag and drop menus, avoiding data entry and harvesting as much information from within files as possible. A beta version is planned for mid-2012.

[17] http://www.lat-mpi.eu/tools/arbil

[18] http://www.language-archives.org/item/oai:paradisec.org.au:External-Raga

[19] http://search.language-archives.org

archives of which fourteen were active (that is, they had material deposited) within the past six months, and nineteen more were active within the past twelve months. This means seven were inactive for the past twelve months. Clearly we need more archives and more that contribute to OLAC.

Once an item has been located it should be available for use if possible (assuming that issues around rights management have been dealt with). Examples are the DOBES data sets, or PARADISEC's online collections of papers by Capell[20], Wurm[21], and Roesler[22]. By late-2012 PARADISEC will provide streaming access to most of its collection.

Accessibility also implies that analog material is digitized. While newly created linguistic records are typically digital, a great deal of legacy material exists only in analog forms and so is outside of the scope of much current language archive infrastructure. For these older materials, we need an effort of discovery and digitization as argued by Schüller, who notes that '80% of the world-wide holdings representing the cultural and linguistic diversity of mankind are not held by audiovisual archives proper' (Schüller 2004: 9), and, further, that analog recordings are in urgent need of digitization if they are to be playable at all.

3. WHAT KINDS OF USES WILL EVOLVE IN THE CONTEXT OF THE SOCIAL MEDIA?

The uses of linguistic data can be online or offline. There are two kinds of online use that need to be distinguished. The first deals with online material as the authoritative archival source. For online use of data there must be persistent location and identification that allow citation and resolution of links, which requires proper repositories with a longterm commitment to curating the material. As can be seen from the figures given earlier, there is too little use of existing language archives (and perhaps a need for more such archives to be established), so, while social media can play a role in dissemination or publicity, without long-term repositories, the data are at risk of loss. Once people start combining data from disparate sources (which could be 'mashups' or could, for example, involve correlating transcripts and media in 'compound objects'[23]), they will create new research objects that themselves may need to be identified and curated in archives (it may be that not all online interactions in small languages necessarily need to be archived in perpetuity).

The second use of online data relates to its use in presentation systems, which may (but need not) be ephemeral. Distinguishing these uses is critical as there are many examples of considerable effort being devoted to presentation systems for community use ('mobilization') which are then lost as the delivery system (which could be proprietary software or websites that are no longer maintained) becomes unusable. If the 'mobilized' material is unique, it poses real problems for longevity, but if it is derived from already archived material it is, essentially, ephemeral.

Offline use is likely to be most relevant to speakers of the languages recorded, given the lack of affordable – or indeed any – internet access. Such offline use of language records includes printed outputs and media on CD, DVD, or in computer-based (e.g. iTunes) formats. The formats in which data are initially created during fieldwork are crucial here too:

---

[20] http://paradisec.org.au/fieldnotes/AC2.htm

[21] http://paradisec.org.au/fieldnotes/SAW2/SAW2.htm

[22] http://paradisec.org.au/fieldnotes/ROES/web/roes.htm

[23] http://www.openarchives.org/ore/

well-formed and predictable data can be readily converted from an archival form to a deliverable form. For example, a dictionary of a language should be derived from a structured lexical data set, as in Toolbox. Similarly a set of texts for production in a book can be derived from a set of interlinear glossed texts in Toolbox. Books can now be produced relatively cheaply in publish-on-demand systems[24], with downloadable versions of the pdf file available via a suitable online repository. Media for a CD or DVD can be readily converted from high-resolution WAV or JPEG2000 files to playable formats for delivery on a CD or used in iTunes installations.

4. CONCLUSION. To conclude, having made some inroads into the production of enduring and reusable records of endangered languages, we still have a long way to go. There is a need for research into existing and emerging methods and development of tools both for creating linguistic data and then for making it useful. There is a need for data management skills to be developed among linguistic scholars so that our relatively small collections can be maintained. We need good descriptive systems (metadata) and simple systems for metadata entry as well as more repositories to hold the material. We all know of projects which have been completed and for which there are now large datasets that are not being properly maintained. The few present language archives are already stretched and cannot go looking for collections, but without such active seeking many collections will be lost.

For those outside of the present training and funding regimes there is a great need for advocacy to promote good practices in working with digital data and to bring to their attention ways of working that will make their work easier and will also have better outcomes for data sharing or reuse. This means more training in both academic and community settings and more sharing of experience and methods (using lists like RNLD, for example). Linguistic archives (e.g. PARADISEC, ELAR, and DoBeS) typically provide advice and support via their webpages and via regular training courses. There needs to be much more activity to allow new researchers to build their own collections and to assist established researchers in describing existing collections. Finally, we need to create incentives for creating the kinds of collections described here. Such incentives include academic recognition of the effort put into building and describing one's research collections and then lodging them in a suitable repository. Citing data from its archival source will also enhance the visibility of collections, and we should now, as authors, reviewers, and editors, encourage the use of such citations in academic papers.

REFERENCES

Aikhenvald, Alexandra Y. 2007. Linguistic fieldwork: Setting the scene. *Sprachtypologie und Universalienforschung - STUF (Special Issue: Focus on Linguistics Fieldwork, ed. Alexandra Y. Aikhenvald)* 60(1). 3–11.

Beagrie, Neil, Julia Chruszcz & Brian Lavoie. 2008. *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities*. London: Higher Education Funding Council for England.

[24] My collection of texts in South Efate was produced in this way and is for sale online as a book here: http://www.bookshop.unimelb.edu.au/cbc/p?IS.9781921775505, or for free download as a pdf file here: http://repository.unimelb.edu.au/10187/9734.

Borgman, Christine L. 2007. *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Dixon, R.M.W. 2006. Dixon Accepts Bloomfield Award. SSILA [The Society for the Study of the Indigenous Languages of the Americas] Newsletter Number 236: 6.

Dobrin, Lise M., Peter K. Austin & David Nathan. 2009. Dying to be counted: the commodification of endangered languages in documentary linguistics. In Peter K. Austin, Oliver Bond & David Nathan (eds.), *Proceedings of Conference on Language Documentation and Linguistic Theory*, 59–68. London: School of Oriental and African Studies. http://www.hrelp.org/publications/ldlt/papers/dobrin_austin_nathan.pdf (21 March, 2012).

Newman, Paul. 1992. Fieldwork and Field Methods in Linguistics. *California Linguistic Notes* 23(2). 3–8.

Newman, Paul. 2005. Field methods courses in linguistics. Paper presented at the Linguistic Society of America conference on Language Documentation: Theory, Practice, and Values. July 9–11, Harvard University.

Newman, Paul. 2009. Fieldwork and field methods in linguistics. *Language Documentation and Conservation* 3(1). 113–125.

Schüller, Dietrich. 2004. Safeguarding the Documentary Heritage of Cultural and Linguistic Diversity. Language Archive Newsletter 1(3), 9–10.

Nick Thieberger
thien@unimelb.edu.au