A balancing act: community interests and language documentation.

Dr Anthony Jukes, Endangered Languages Academic Programme, SOAS, London

One of the central goals of language documentation (as laid out by Himmelmann (1998) and others) is the creation of corpora of language data — usually consisting of a core of digital video and audio recordings and associated annotations and metadata. For a variety of reasons (including security and convenience, but also as a requirement of funding agencies), these corpora are increasingly being housed in dedicated digital endangered language archives such as PARADISEC in Australia, ELAR in Britain, and DoBeS in the Netherlands.

However, it can be hard to explain the value of such digital archives to communities in the developing world, often without access to computers and high-speed internet. Rather, the practice of collecting material with the express purpose of sequestering it in first-world institutions can appear (at best) irrelevant, and (at worst) a throwback to colonial practices of cultural expropriation. Showing the value of documentation in general and archives in particular can be a challenge for fieldworkers and archivists, which requires flexibility and a willingness to discover and adapt to community goals. This paper will discuss issues that arose during a documentation project in North Sulawesi, Indonesia, and will show some examples of material that came out of the project.

Language documentation

Recent years have seen the emergence of a new field of linguistics: language documentation, which is ‘concerned with the methods, tools, and theoretical underpinnings for compiling a representative and lasting multipurpose record of a natural language or one of its varieties’ (Gippert, Himmelmann, & Mosel, 2006, p. v). Language documentation differs from language description (which aims to describe a language's system of structures and rules in the form of a grammar or dictionary) because of its focus on primary data—though of course documentation does not preclude description and is in fact an excellent corollary to it.¹

Language documentation is most often associated with work on endangered languages, and it is fair to say that an increased interest in language endangerment (both among linguists and in the wider community) was the primary motivation behind its inception (and likewise the rationale behind the major funding schemes), but the principles behind it can be applied to any linguistic work involving use of primary data.

The most visible ways in which language documentation is changing the field are: a large and growing literature, the creation of specialised training courses for linguists and language workers, the funding of a significant number of documentation projects

¹ It is also the case that most projects of language description necessarily involve the collection of large amounts of primary data. It is the contention of documentary linguists that this data should be archived or otherwise preserved for posterity and further use by the speech community or its descendants, by linguists (eg. so theoretical claims about the language can be checked against the data), or for other unforeseen uses.
around the world, and the associated emergence of specialised digital archives for endangered language material. It is the latter two phenomena that I wish to discuss here.

**Documentation projects**

The emergence of language documentation has seen the creation of a new type of linguistic research activity: the dedicated language documentation project. These have been aided by the availability of funds specifically for this purpose. Since the requirements of funding agencies are important in shaping the form of projects it is worth listing the major ones here. The current main sources of funding for academic language documentation projects are:

- The DOBES programme funded by the Volkswagen Foundation, begun in 2000 and still ongoing. It has funded 30 documentation teams and has also been prominent in the development of software tools and both practical and theoretical standards of documentation.
- The Endangered Languages Documentation Programme (ELDP), funded by Arcadia. Since 2003 ELDP has funded projects on around 130 languages, and generally allocates around £1 million in grants (out of a bequest of around £17 million) per year.
- Documenting Endangered Languages (DEL) funded by the National Science Foundation (NSF) and the National Endowment for the Humanities (NEH) in the USA. Initiated in 2002, it aims to award $2 million annually.

There are other, smaller, sources of funds, and it appears that as documentation becomes more recognised as a field there will be more projects funded by the usual academic funding bodies such as the Australian Research Council or the Economic and Social Research Council in the UK.

**Typical outcomes of documentation projects**

Although the field is in its infancy and there is much debate about what a ‘proper’ documentation might look like, the description from the DoBeS website is illustrative:

The documentation, and therefore the archive, contain the following types of material:

- annotated audio and video recordings of diverse speech events with transcriptions, translations into one or more major languages, morphosyntactic analysis and other comments on content and linguistic phenomena
- photographs and drawings partly bundled into groups of photos documenting processes, as, how to build a house
- music recordings and videos of cultural activities and ceremonies
- a description of the language's genetic affiliation, its socio-linguistic context, its phonetic and grammatical features, and the circumstances of research, recording and documentation

---

2 Again, linguists (and others) have always been involved in collecting language data. What is new is the emphasis on data collection as a central activity.
• keyword-based descriptions to facilitate the organization and accessibility of documents in the archive

ELDP adds a grammar, dictionary and thesaurus to the list of desirable contents of a documentation. But clearly, the core of the documentation project is seen to be the corpus of annotated recordings.

To give an example of what this might look like, consider the documentation project of Toratán, a language spoken by about 150 people in three villages in North Sulawesi, Indonesia. The main outcome of this project was about 50 hours of recordings (mostly video), comprising 70 individual sessions. The bulk of the sessions are of a genre which, given the advanced age of the speakers, could aptly be called 'extended conversational reminiscence', though there are also more straightforward procedural narratives (how to make palm sugar, how to prepare particular foods, the proper way to conduct a wedding, etc), and personal stories. Some are of a single speaker, but the majority are of a pair or trio of speakers. In total 20 speakers feature in the recordings, varying in age from 45 to 86.

The sessions were all transcribed and translated into Manado Malay (the local variety of Indonesian) and English. The annotations were prepared using the ELAN (EUDICO Linguistic Annotator) program developed for DoBeS by the Max Planck Institute in Nijmegen. The ELAN file is an xml file containing information about the media files, segmentation in the form of timecode, and annotations which are linked via those timecodes to the media files. A subset of recordings were further analysed using Toolbox resulting in a further set of interlinear tiers.

Figure 1 shows the ELAN program open with a recorded session (a discussion about the sequence of a traditional courtship and wedding) and its annotations. There are tiers for transcription, translation into English and Manado Malay, and interlinear text showing word, morpheme, gloss and part-of-speech. Not shown here is the associated metadata about the date, type, and circumstances of recording, speaker identity, genre and so forth.

---

4 This was a two year ELDP-funded post-doctoral project (IPF0087) carried out by the author.
Figure 1: A video recording and associated annotations in ELAN

An annotated recording such as this lends itself to a variety of uses for the linguist, in that particular recordings (or indeed the whole corpus) may be searched for instances of particular words or morphemes, which may then be seen and heard in context. Depending on the level of annotation, there may also be information about gesture, utterance type, constituent order or whatever other sorts of information are found to be important.

The above is simply given as a representative example of the type of data constituting the core of a language documentation. There is of course variation across projects. Some have a particular focus, eg. ethnobotany or marine science. Some have made many more hours of recordings, but annotated only a small proportion. Some have audio rather than video recordings. Many use different tools for transcription and analysis. But they are all repositories of primary language data, with added information to make it searchable and understandable.

It is fair to say that the central goal of a documentation project as funded by any of these three bodies is to collect, annotate, and archive language data of this type. This leads fairly naturally to a widespread assumption among documenters that the ‘purpose of their work is to “deposit data in the archive”’ (Nathan & Fang, 2008). Thus, the researchers give the outcomes of their documentation project to the archive,
provide information to be used in cataloguing and the documentation is in some sense considered to be ‘finished’. The archive will look after the material from then on, in perpetuity, and allow access to it according to the requirements as set by the researcher and/or the community.

**Endangered language archives**

Accompanying the emergence of language documentation has been the need for digital archiving of endangered language materials, and the creation of dedicated archives for this purpose. ELDP and DOBES explicitly require all material to be deposited in their own digital archives. DEL similarly requires digital archiving of data but does not specify the destination, merely advising grantees to conform to ‘best practice’ as recommended by E-MELD (Electronic Metastructure for Endangered Languages Data) and thus to deposit materials with an archive belonging to OLAC (the Open Language Archives Community).\(^5\)

The mission statement of the Endangered Languages Archive (ELAR), based at SOAS in London, is as follows:

We aim to:

- provide a safe long-term repository of language materials
- enable people to see what documentation has been created for a language
- encourage international co-operation between researchers
- encourage endangered language communities to participate and to build on the work done, in order to safeguard their languages
- provide advice and collaboration\(^6\)

Access to the contents of the archive is generally via an online catalogue. To illustrate, the following screenshot shows the DoBeS archive, with metadata about a particular recording in the Saliba/Logea project headed by Anna Margetts at Monash University. Subject on access restrictions, a user may search the archive, and via the catalogue/browser download media and annotation files, or view/hear streaming media.

---

\(^5\) OLAC has 37 participating archives, relevant ones here include ASEDIA (Aboriginal Studies Electronic Data Archive), AILLA (Archive of the Indigenous Languages of Latin America), and PARADISEC (Pacific And Regional Archive for Digital Sources in Endangered Cultures).

\(^6\) [http://www.hrepl.org/archive/](http://www.hrepl.org/archive/)
Figure 2: The DoBeS catalogue

**Documentation, archiving and community interests**

The nature of the relationship between a fieldworker and the host community has been the subject of considerable debate within linguistics and other disciplines. It has grown from the simple notion that a researcher should ‘give something back’ to the community to the argument that a fieldworker should actively work for the community, or that the fieldwork should be considered an entirely collaborative project — fieldwork by a community (Grinevald, 2003). How this might best be achieved is a matter of considerable debate, and in any case the needs and desires of language communities can differ in many ways. Sometimes they may be hard to gauge as the community itself is not necessarily clear or undivided about the potential uses of a linguist.

It has been my experience (and that of others, such as Musgrave and Thieberger (2007)), that while communities are generally not opposed to documentation work, it is unlikely be a priority for most people.⁷ Interest in the language is often low — this is usually one of the reasons the language is endangered, thus it is difficult to claim in any real sense that the fieldwork is being carried out by the community. And I can

---

⁷ These observations are based on fieldwork in Indonesia and the South Pacific. I readily concede that they are not universally applicable.
only agree with Musgrave and Thieberg when they state that ‘if such communities could access the money [ie. from ELDP etc], they would not in general choose to spend it on supporting language documentation’. If they were to spend money at all on language matters it would probably be on schoolbooks as part of a revitalisation programme: however two of the funding bodies (ELDP and DEL) explicitly state that they will not fund projects primarily concerned with revitalisation. (DOBES does not state this, but implies it by omission.)

Furthermore, the end result of a standard documentation project — the archived material — is likely to be inaccessible to communities in the developing world. In the Toratán case, although North Sulawesi has reasonably good communication infrastructure, computer ownership and access is very low, and the internet is virtually unknown outside of the provincial capital Manado (a gruelling 4 hour trip from the Toratán villages). Broadband internet access is not widely available, and even where it is, its price is prohibitive and is likely to remain that way for some time. Thus, online access to material archived at ELAR, PARADISEC or DoBeS is not currently feasible. The material can be stored on a local computer, but low levels of computer literacy and the complexities of the documentation software make this a poor solution.

In addition, it has been pointed out that although the rhetoric of endangered language appeals to the irreplaceable cultural and environmental knowledge held within the language, ‘studies carried out by linguists tend to focus on structural and lexical analysis’ (Batibo, 2005, p. 40). Even when there is culturally important information contained within documentary material, the currently defined metadata schemes such as IMDI and OLAC emphasise the research interests of typological linguistics rather than aiming to be ‘multi-purpose’ in the sense of Himmelmann (2006, p. 2), ‘even though there is no evidence anecdotally or in the literature that typological findings are valued by or useful for language speakers’ (Nathan & Fang, 2008). Thus, even if the archived material were accessible, it is unlikely in that form to be of much use or interest to a community.

The above considerations are undeniably depressing, and lead to the question as to whether documentation can help speech communities at all. The good news is that there is a healthy and ongoing debate about these matters within the community of language documenters. The archivists have also shown themselves to be aware of these problems, and are attempting to ensure that materials are accessible to language communities and the nations in which they reside. PARADISEC, for example, has an agreement with cultural institutions in three countries (Vanuatu, New Caledonia, and Papua New Guinea) ensuring that digital materials are made available in accessible forms (Nick Thieberger, pers.comm). DoBeS has a policy of setting up regional archives in order to improve access for local researchers and language workers. Meanwhile, the archivist at ELAR has written and published extensively on the need for archives to be involved in ‘mobilisation’ of material (Nathan, 2006; Nathan & Csató, 2006), and (with others) has heavily criticised the emerging orthodoxy of documentation and the ‘commodification’ of endangered language research (Dobrin, Austin, & Nathan, 2007).

For the Toratán case, I simply decided to make most of the archival material available to the speakers in a readily accessible form. Since most of that material was video, and since speakers I worked with enjoyed watching the recorded conversations, I gave
DVDs to speakers and family members. These needed to be subtitled in Manado Malay if the majority of community members were to understand, as only those over 60 are fluent speakers. Conveniently, the material I had prepared for documentation already had annotations in the form of transcriptions and translations that could be used to make these subtitles. Although not an especially straightforward process (involving various methods of text manipulation and the use of a DVD authoring program), it was possible to produce these on my laptop in the field. The end result is DVDs which can be played on any consumer player. The subtitles default to Manado Malay, but can be changed via the DVD remote control. A screenshot showing how one of the procedural texts (how to make palm sugar) looks with the English subtitle appears in Figure 3.

![DVD with subtitle](image)

**Figure 3: DVD with subtitle**

People enjoyed watching them, including children, who generally stopped whatever they were doing to gather round and watch. My main consultant, Bert Hosang (pictured at right in Figure 4), reported that this was the first time in many decades that a child had shown any interest in Toratán.
Conclusions

Language documentation is a field in its early stages, and is continually evolving. We should be aware of problems and imperfections but we should not allow them to distract us from the central issue: while debates about the proper methodology and ethics of documentation unfold, languages continue to fall silent. While it would certainly be preferable for all documentation projects to satisfy all potential users, it is nevertheless the case that even imperfect documentation is better than none, and also that the documentation projects taking place today are providing a far better record of endangered languages than anything carried out in the past. As the field matures, this should only improve.

References


