Documentation in practice: Developing a linked media corpus of South Efate

Nicholas Thieberger

1 Abstract[1]

There is a growing need for linguists working with endangered languages to be able to provide documentation of those languages that will serve two functions, not only the analysis and presentation of examples and texts, but also the means for accessing the material in the future. In this paper I describe a workflow for building documentation into a language description developed in the course of writing a grammar of South Efate, an Oceanic language of Vanuatu, for a PhD dissertation. I suggest that, with appropriate tools, the effort of recording and transcribing documentary field recordings can result in a media corpus from which we can produce instant links between text and media, which in turn enriches our analysis. Further, these annotations are in an ideal form for archiving and for providing access to data by the speakers of the language. I take it as axiomatic that we must archive our recordings and associated material and that this step is integral to the larger project of language documentation.

2 Documentary output

Linguists working on small and endangered languages are being exhorted to produce their data in reusable forms (see for example Bird and Simons 2003) and at the same time to increase the scope of the recorded material so as to document as much as possible of the language and the knowledge encapsulated in the language (see Himmelmann 1998, and Woodbury 2003). These two goals are ambitious, and especially so if the language recording is part of a PhD programme. An acceptable and normal PhD dissertation focussed on an indigenous language requires a grammar that covers the traditional components of phonology, morphology and syntax. A typical language description should also include what we might call a 'Boasian trio' of grammar, texts and a dictionary which represent an analytic apparatus built on the corpus. The corpus may also include audio and video recordings and photographs that are linked to the analytical apparatus of the grammar. With analogue material it has not been easy to build an accessible corpus and it was rarely attempted in the past. While such a corpus may be referenced in the grammar, for example, it is not accessible beyond the physical location in which it is stored, typically a library in a metropolitan centre. Digital data, on the other hand, can be accessed instantly and, if archived, becomes a citable resource into the future.

Managing all of the relationships between parts of a linguistic dataset can quickly become overwhelming, but with the use of appropriate tools we can increase our access to the

[1] Thanks to Birgit Hellwig and David Nathan for comments on an earlier draft of this paper.

data and maintain many of its internal relationships. The results of this work then conform to those desiderata broadly labelled as language documentation (discussed further in §3 below). In short then the question posed in this paper is how can we extend the documentary aspect of our work so that the primary recordings, and our annotation of them, become part of our output?

While multimedia techniques are increasingly being used to represent ethnographic and linguistic information, they are rarely implemented in a way that provides long term benefits and thus are not ideal documentary outputs. A main reason is that they are not produced by linguists who are aware of the underlying issues surrounding proper archiving and long term access to field materials. The data is typically heavily edited, and does not represent a primary and citable source, both of which are commonly accepted as central tenets of good documentary practice (cf Bird and Simons 2003 and the discussion below). Ideally, each segmented piece of data in the multimedia package needs to be citable and related to an archival object – which is not the case, in my experience, in any current multimedia language package[2]. Furthermore, these 'packages' are typically in a proprietary format that does not allow the data to be reused. As standalone objects, delivered on a CD, they do not facilitate remote access to media objects, although this could be overcome by making copies available over the web. 'Multimedia' should be derived from well-formed media and should not be the sole representation of the data.

To build well-formed data we need a method for working with archival data, or with a derived form that has most of the characteristics of the archival form. Thus, for example, an archival audio file can be downsampled to a manageable size using linear MP3 compression, but still maintain offset points or timecodes relative to the start of the file (working with MP3 versions means that large numbers of files can be located on a single hard disk for instant access). A method for working with archival media is discussed in the following section.

3 Practicalities of data linkage

In my PhD dissertation, a grammatical description of a language of Vanuatu, I wanted to provide source information for each example sentence and text that would allow the reader to locate the example within the field recordings so as to be able to verify that the example actually did occur in the data. I wanted the field tapes (both audio and video) themselves to be accessible with sufficient descriptive material (including simple metadata) to be locatable and citable. Being citable entailed a persistent identification and location for the data in a trusted archival repository. To do these relatively simple tasks it was first necessary to link the transcripts of the field tapes to their audio source. In 1998 when I began doing this work, there were principles and methods pointing in the direction of reusability of linguistic data (as later formulated by Bird and Simons 2003), but the tools to do the work were either nascent or non-existent. By building my own tool called *Audiamus*[3] and conforming to these

[2] But Csató and Nathan (2003:81) would suggest they are an exception.
[3] On *Audiamus*, see: http://www.linguistics.unimelb.edu.au/thieberger/audiamus.htm

principles I have been able to capitalise on the general linguistic community's development of tools, for example for aligning a transcript to a digital sound file (such as *Transcriber*[4], for example, which can take the output of *Audiamus*). The thesis was presented together with a DVD of some 3.6 Gbytes of data, representing over 18 hours of transcribed and linked media data and presented in a cross-platform stand-alone version of *Audiamus*.

For some years now I have observed practitioners using computers for language work of various kinds and, in general, we use what we are most familiar with and what is easiest for us to incorporate into our normal workflow. While this need not be 'bad practice', the crucial point is that the underlying principles of reusability and interoperability of the data are observed regardless of what tools are employed. In the absence of a concerted training effort, we must avoid purist dictates about what is **the** correct way to proceed and to encourage appropriate use of existing tools until we have purpose-built tools that are generally used and accessible to ordinary working linguists. Presentation of the data in an ideal form carries with it certain implications for the data which are listed below (adapted from Bird and Simons 2003) together with an outline of the approach taken in my PhD thesis.

- The data is stored on media that will persist into the future; *All field tapes have been digitised and are stored in the archive established by PARADISEC[5] with copies held at both the University of Sydney and at the Australian Partnership for Advanced Computing*.
- The data is adequately described using standard controlled vocabularies so that it can be located; *The metadata, or cataloguing information associated with the deposited material conforms to the present standards agreed to by the Open Language Archives Community (OLAC[6]) and uses their controlled vocabularies for the role of participants, language names, and so on.*
- A description of the data is available for researchers via standard search mechanisms; *The metadata can be searched via OLAC, or on the LinguistList[7] pages. As the metadata conforms to the Open Archives Initiative (OAI)[8] guidelines it can also be searched by any OAI conformant search engine.*
- The data is in a form that will be legible over time (not locked into transient proprietary formats) and documents the use of any special fonts; *The transcripts of audio files are in plain text format marked-up to show the timecodes that relate to the media files. The dictionary of South Efate is also a plain text file that is exported via dedicated lexicographic software (Shoebox) to a formatted rtf file. The special characters required for South Efate are described in the metadata, and are rendered as m$ and p$ in ASCII format, the equivalent forms required for representation using the IpaTimes font. Future work will render them using Unicode.*

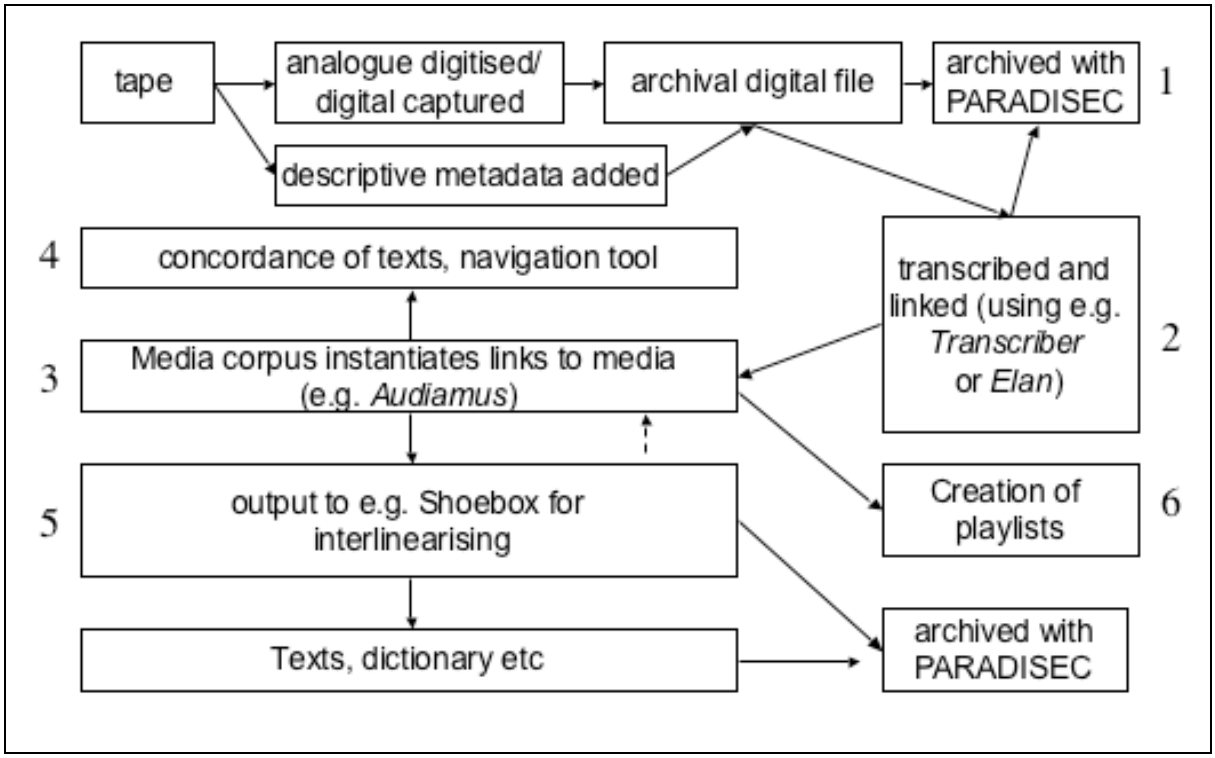
[4] http://www.etca.fr/CTA/gip/Projets/Transcriber/
[5]Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) see http:paradisec.org.au
[6] http://www.language-archives.org
[7] http://linguistlist.org/olac/
[8] http://www.openarchives.org/

- The presentation and the structure of the data are kept separate so that the former is derived from the latter; *This is the most crucial point for working with digital data. Presentation formats for the South Efate data derive from its structure (for the texts, transcripts and dictionary). The grammar is produced in a word-processor and will need to be converted to a more suitable format when one becomes apparent. It is archived in pdf format.*
- Recordings are provided unsegmented and with time-aligned transcripts to allow others to verify an analysis; *Recordings are archived and reference is made to the unsegmented audio file to maintain the context of utterances and to allow other researchers to access the data*.
- The data is described at a level of granularity that allows citation of individual utterances; *Audio files are transcribed to utterance level, and, using* Audiamus*, are citable to that level so that example sentences in derived work can be given persistent citation forms.*
- Copyright and intellectual property conditions are explicit and enforced; *Copyright and moral rights of speakers are asserted in the present work, and the archived data can only be accessed by password. Speakers have a contract that specifies what uses their recordings can be put to*.

4 Digitisation of the audio file, first step in the wworkflow

In the next part of this discussion, I outline the workflow developed for working with a digital media corpus using *Audiamus* (Figure 1), showing that the input is a linked text file and a digital media file and the textual outputs can be in several formats. Once we have digitised the media file and provided descriptive metadata to allow it to be discovered it can be archived (cf 1 in Figure 1). This provides us with a citable archival file on which we can base our analysis. Its identification in an archive will persist, so our references to the data in this form should be available to users in the distant future. Such references are the transcript and subsequently derived examples that we will use in our grammatical description.

Figure 1: Workflow using Audiamus



I conducted fieldwork on the language of South Efate in Central Vanuatu in several fieldtrips from 1996 to 2000. The resulting field tapes contained monologic narratives, conversations and court hearings and were recorded on analogue tape and some digital video with a range of male and female speakers of different ages.

On my return from fieldwork I began digitising my analogue tapes using the built-in soundcard of a desktop computer. This was a mistake! I ended up with digital audio files that I then used to align with the transcript. However, these were not good quality audio files because the computer's soundcard is simply not adequate to the task. Thus, when the opportunity arose to have the analogue tapes digitised at a higher, and archival, resolution it resulted in my having two versions of the digital data. These two digital versions of the same tape did not correspond in length due both to stretching of the audio tape, and to being played on different cassette players, with slightly different playback speeds. There was no simple correlation between the timecodes in the old and the archival version. While I linked all subsequent transcripts to the archival version of the audio file, due to the time constraints of dissertation writing I have kept the non-archival versions for presentation of the thesis data. Archival versions have been lodged with PARADISEC[9].

The crucial lesson from this experience is to digitise field tapes at the best (archival) resolution possible and then use those files (or a down-sampled version, such as MP3[10] if the

[9] Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), http://paradisec.org.au

[10] MP3 files can be indexed by timecodes as long as the mp3 files are not encoded with a variable bitrate. Note that MP3 is absolutely not a suitable format for recording or archiving.

original is too large), as the basis for linking to transcripts. To produce the best quality digitisation, it is recommended to use external soundcards that avoid computer noise, that is, if you don't have a friendly digitisation project at a campus near you. It is also worth keeping up with the technology[11] to find out about new methods and media for doing this kind of work.

5 Linked transcription of the digital media

All my field tapes were transcribed, mainly by a speaker of South Efate using a ghetto blaster. He wrote the transcripts in exercise books with translations in Bislama, the national language of Vanuatu. These South Efate transcripts were then typed and imported into text/audio alignment software (cf 2 in Figure 1). When I began doing this work in 1998 I used Michel Jacobson's *SoundIndex*[12] from CNRS/LACITO to align the transcript to the media file (the current tool is Claude Barras' *Transcriber*).

The transcribing software produces a number of different outputs, among them a simple text file which has the utterance chunk together with the start time and end time in the audio file. These linked transcripts are an index of the content of the field tape and can be archived together with the media file. For the purposes of analysis of the data, the links need to be instantiated, that is, we need to be able to click on a sentence and hear it. Using the transcription tool it is possible to do this for each individual file which then has to be opened and searched individually for any given form. However, I needed a corpus of all transcripts, with a concordance showing all word forms. There was no such tool available at the time so I used HyperCard[13] to construct the links in a way that allowed the data to be imported and exported easily, using Quicktime to instantiate the links to offset points within large data files. This tool is called *Audiamus* (cf 3 in Figure 1).

Using HyperCard may appear to be a retrograde step, but it capitalised on my existing knowledge of the software, and also allowed me to use a well-developed concordancing tool written by Mark Zimmerman, called *Free Text*[14] (cf 4 in Figure 1). Combining these tools resulted in a keyword-in-context concordance of texts that played the audio of the context of the selected items (typically the context sentences). I developed the HyperCard tool over several years to provide access to the media for the purposes of linguistic analysis. The audio file (in .wav or derived MP3 format) is addressed as an external object and the text and timecodes are routinely exported to plain text files for storage, thus ensuring that the data can flow through the tool rather than being incorporated into it and thus potentially locked away from future means of reading it. It is useful to view the data as a stream that flows through the various tools that we use to analyse it, and to be aware that the stream must flow out of a tool as easily as it flows into it.

---

[11] See for example the links on digitisation here: http://www.paradisec.org.au/Digital_Links.htm
[12] http://lacito.vjf.cnrs.fr/archivage/index.html.fr
[13] In 2002, *Audiamus* was rewritten (as version 2) in Runtime Revolution, a cross-platform application that can build standalone distribution versions of *Audiamus*.
[14] http://www.his.com/~z/c/index.html

Audiamus is designed with the key principles of reusability of and accessibility to the data, with the basic premise that every example quoted in the thesis should be provenanced to an archival source if possible. The importance of a simple tool like Audiamus is that it provides a means for working interactively with field recordings in a way that was not otherwise possible. In Audiamus the user can select an example and clip either its timecodes or both the text and the timecodes to the clipboard for pasting into a document. The timecodes are specified as follows: (audio filename, start time, end time) or (98002b, 1413.9999, 1419.3600).

Examples can also be clipped with timecodes in a format suitable for processing in Shoebox, e.g.:

\aud 98002b

\as 1413.9999

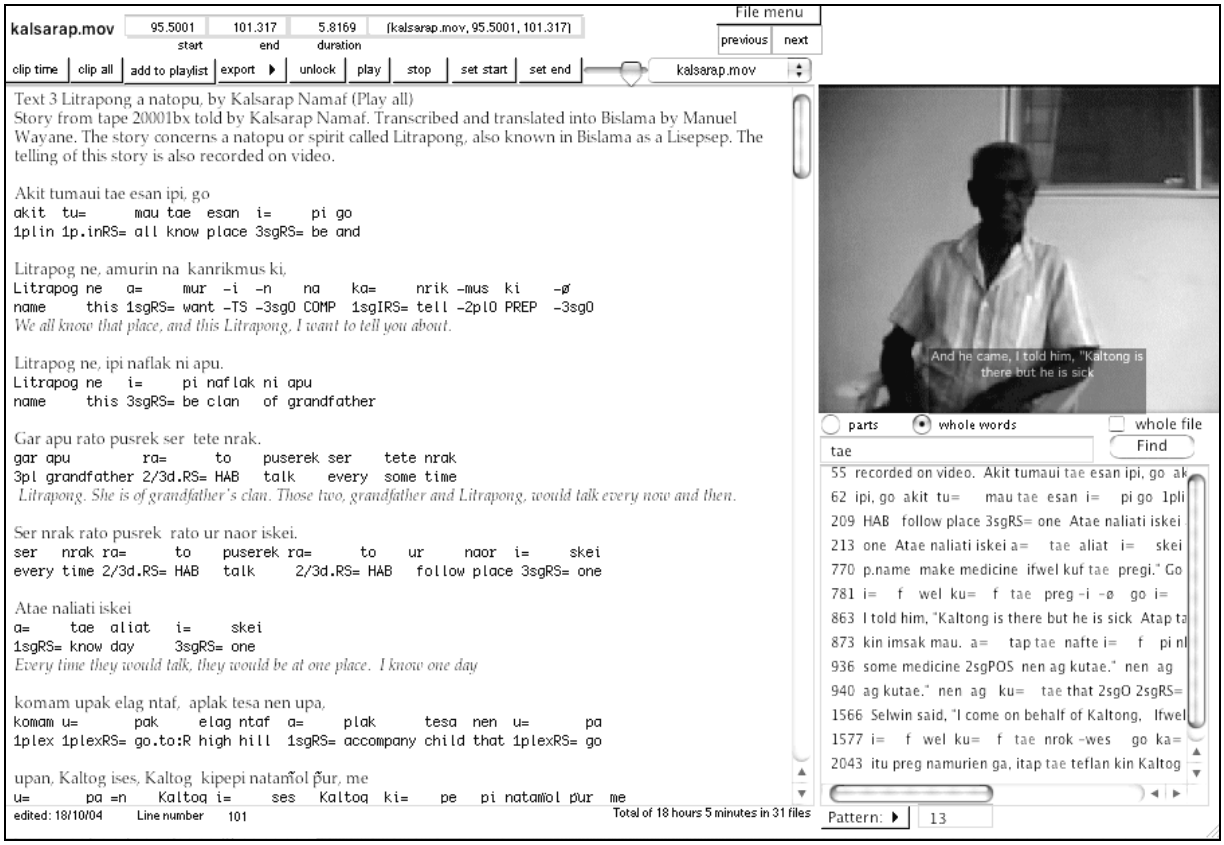
\ae 1419.3600

\tx (line of text)

The time-aligned text is then available for interlinearising in Shoebox while maintaining the timecodes (cf 5 in Figure 1). The output of this process can then be re-imported into Audiamus so that the interlinearised version becomes playable, as illustrated in Figure 2.

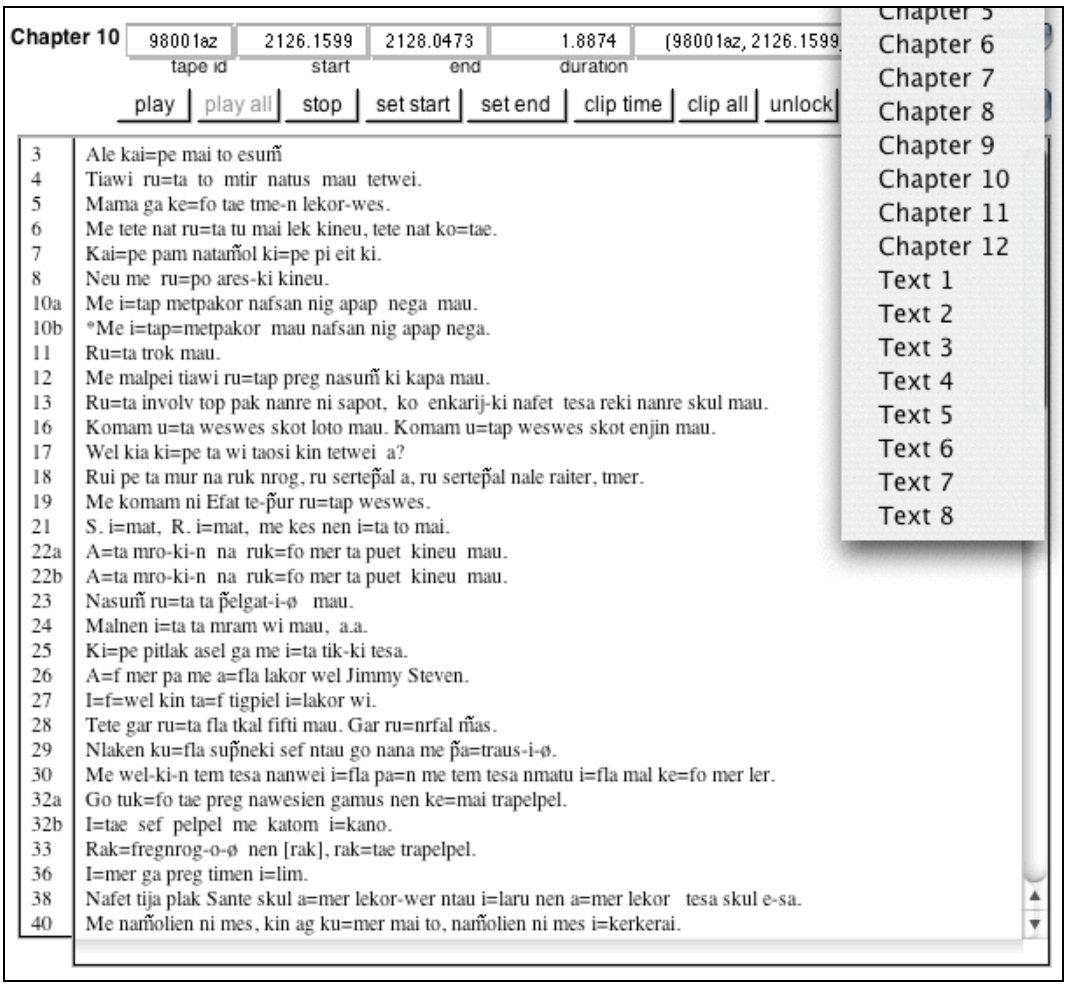
Another export routine provides Quicktime text tracks with timecodes in the appropriate format (hr:min:sec:frame) which can be used as subtitles to Quicktime movies, as seen in the image in the right hand top corner of Figure 2. In this way, it is possible to create video subtitles in either the local or the metropolitan language.

Figure 2, Screenshot of Audiamus showing interlinearised time-aligned text. The search frame in the bottom right lists all occurrences of the search string and a mouseclick on any line plays the corresponding media.



Examples selected from the corpus can be added to a playlist (cf 6 in Figure 1) to illustrate particular constructions, no matter where in the data they occur. The playlist itself can be stored for future use while another playlist is constructed. These playlists can then be used in presentations, seminars or talks to play illustrative examples. In the presentation of my dissertation, each chapter was a playlist consisting of numbered examples (shown in Figure 3), allowing readers to hear the sentences quoted directly from the archival media. These playlists offer the ability to recombine the index of the data from its original version as a transcript of an entire media file to a selection of items chosen to facilitate comparisons. The playlists leave the data itself intact so that contextual information is still available to the user.

Figure 3, Playlist function of Audiamus. The screen in this example is a chapter of my thesis. Each chapter and example text is listed in the popup menu on the right. Examples are numbered (in the leftmost column) as in the thesis.



6 Conclusion

Language documentation necessarily includes the production of archival data. By using Audiamus and adopting the principle of data reusability, it has been possible to produce an archivable, citable, extensible set of data and to construct links between my field recordings and the grammar, thus presenting both the analysis and the primary data for verification.

If we build data linkage into our workflow as part of normal linguistic analysis, we end up with richer descriptions based on contextualised and verifiable data which has more archival use than does a set of media files or cassette tapes alone. The shift in our relationship to the data will be far-reaching as it entails a shift in the authority of the analysis away from the analyst and to the data.

7 References

Bird, Steven and Gary Simons. 2003. Seven Dimensions of Portability for Language Documentation and Description. *Language* 79: 557-82

Csató, Eva and David Nathan. 2003. Multimedia and documentation of endangered languages. In Peter K. Austin (ed.), *Language documentation and description*, Volume 1, 73-84. London: SOAS.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36: 161-195

Woodbury, Anthony C. 2003. Defining documentary linguistics. In Peter K. Austin (ed.), *Language documentation and description*, Volume 1, 35-51. London: SOAS.

Author/s:
Thieberger, N

Title:
Documentation in practice: developing a linked media corpus of South Efate

Date:
2004

Citation:
Thieberger, N. (2004). Documentation in practice: developing a linked media corpus of South Efatein. In P. Austin (Ed.), Language documentation and description, vol.2, (pp. 169-178). Hans Rausing Endangered Languages Project, School of Oriental and African Studies, University of London.

Publication Status:
Published

Persistent Link:
http://hdl.handle.net/11343/34484