Part I

Documentation

1 Language documentation and meta-documentation

Peter K. Austin

1 Introduction

The past fifteen years have seen the emergence of a new sub-field of linguistics that has been termed 'language documentation' or 'documentary linguistics' (Himmelmann 1998, 2002, 2006, Lehmann 2001, P. Austin 2010a, Grenoble 2010, Woodbury 2003, 2011a). Its major goal is the creation of lasting multi-purpose records of languages or linguistic practices through audio and video recording of speakers and signers, and annotation, translation, preservation, and distribution of the resulting materials. It is by its nature multi-disciplinary and draws on theoretical concepts and methods from linguistics, ethnography, folklore studies, psychology, information and library science, archiving and museum studies, digital humanities, media and recording arts, pedagogy, ethics, and other research areas.

The term 'language documentation' historically has been used in linguistics to refer to the creation of grammars, dictionaries, and text collections for undescribed languages (the so-called 'Boasian trilogy'; for discussion see Woodbury 2011a: 163). However, work defining language documentation as a distinct sub-field of linguistics emerged around 1995 as a response to the crisis facing the world's endangered languages, about half of which might disappear in the twenty-first century (the crisis was identified and popularized in such publications as Robins and Uhlenbeck 1991, Hale et al. 1992, Wurm 2001). Linguists drew attention to an urgent need to record and analyse language materials and speakers' linguistic knowledge while these languages (or threatened special registers and varieties within them) continued to be spoken, and to work with communities on supporting threatened languages before

Versions of this paper have been presented as talks at the Kioloa Aboriginal Languages Workshop 2010, the Linguistic Society of America Annual Meeting 2011, the International Conference on Language Documentation and Conservation 2011, and the seventh European Australianists Workshop 2012, as well as in classes at SOAS and Tokyo University of Foreign Studies. I am grateful to Lisa Conathan, Lise Dobrin, Andrew Garrett, Geoff Good, Heidi Johnson, Anthony Jukes, Stuart McGill, David Nathan, Julia Sallabank, and Tony Woodbury for discussion of the ideas included; I alone am responsible for the material presented here.

opportunities to do so became reduced. The emergence of language documentation was also prompted by developments in information, media, communication, and archiving technologies which make possible the collection, analysis, preservation, and dissemination of documentary records in ways which were not feasible previously. In addition, it was facilitated by large levels of research funding support from three main sources: the DoBeS (Dokumentation Bedrohter Sprachen 'Documentation of Endangered Languages') programme sponsored by the Volkswagen Foundation in Germany (2000–13), the Endangered Languages Documentation Project (ELDP) supported by the Arcadia Trust in the United Kingdom (2002–16), and the Documenting Endangered Languages (DEL) interagency initiative of the United States National Science Foundation and the National Endowment of the Humanities (2005 onwards).

Language documentation concerns itself with principles and methods for the recording and analysis of primary language and cultural materials, and metadata about them, in ways that are transparent and accountable, and that can be archived and disseminated for current and future generations to use. Some researchers have emphasized standardization of data/metadata and analysis and 'best practices' (e.g. E-MELD, OLAC), while others have argued for a diversity of approaches which recognize the unique and particular social, cultural, and linguistic contexts within which individual languages are used (see Dobrin, Austin, and Nathan 2009, Dobrin and Berson 2011).

This chapter is concerned with the role of metadata in language documentation and argues for a broad approach to observation and documentation of the methods, processes, and outcomes of language documentation projects, which we refer to as 'meta-documentation' (or 'meta-documentary linguistics'). It argues that a theory of meta-documentation does not (yet) exist and discusses some techniques that could be adopted for developing such a theory, as well as proposing some of the components that may make it up. The need for fuller reflexivity on the part of language documenters, and linguists more generally, is particularly emphasized.

2 Language documentation (or documentary linguistics)

Language documentation (or 'documentary linguistics') is defined by Himmelmann (2006: v) as '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'. A similar definition is given by Woodbury (2011a: 159) as 'the creation, annotation, preservation, and dissemination of transparent records of a language'. Woodbury (2011a: 161) also notes that the term 'language documentation' is used in another sense, as well namely the outcomes of documenting languages, but proposes to clarify the terminology: 'The sets of records, coherent or not, are often called LANGUAGE

DOCUMENTATIONS; but since that is what we are calling the activity as a whole, I will call such sets LANGUAGE DOCUMENTARY CORPORA (or just CORPORA).' The form and uses of documentary corpora have been explored somewhat by linguists (e.g. within the DoBeS programme – see DoBeS 2005) and, as Dobrin and Berson (2011: 188) argue:

> It does seem clear that documentary linguists have been on relatively comfortable ground in thinking about the PRODUCTS of linguistic research: conceptually distinguishing an annotated corpus or documentation of a language from a higher order description of its patterning . . . reasserting the intellectual value of vocabulary . . . and oral discourse (as represented in texts) alongside grammar, extending the range of documentary outputs to include items like primers and orthographies that are targeted directly at non-academic audiences . . . They have also enriched the inventory of digital data models, formats, and software tools that facilitate documentary research and enable the preservation and dissemination of its results.

There has been rather less discussion about what Woodbury (2011a: 161) calls 'corpus theorization': 'I will call the ideas according to which a corpus is said to cohere or "add up" its (CORPUS) THEORIZATION. Corpus theorizations, and even principles for corpus theorization, can both offer a space for invention and become a matter of contention and debate.' Corpus theorization has been only weakly developed within documentary linguistics. Seifart (2008) attempts to address some aspects of corpus theorization, namely representativeness and sampling, and Lüpke (2010) discusses data collection methods, but few scholars have been explicit about *why* and *how* they are collecting and organizing their particular corpora, other than for some vague notion of 'documenting the language' or 'saving the data'.

In addition to corpus theorization, Woodbury (2011a: 161) also mentions wider issues of what he calls 'project design':

> Of special interest is the range of concerted, programmed documentary activities motivated by impending language loss and aimed at creating a final record. These activities . . . raise questions about the participants, their purposes, and the various stakeholders in the activity or program of activity or project: we may refer to this set of questions as the PROJECT DESIGN . . . of a language documentation activity.

We see corpus theorization and project design as part of meta-documentation, which we explore and elaborate in Section 3. For some speculations about a possible typology of project designs see Section 4.

3 Meta-documentation (or meta-documentary linguistics)

From the outset, those who have been working on language documentation have been clear that alongside collecting and analysing data (typically audio

and video recordings, but also still images)[1] it is necessary to record and analyse metadata, data about the data, to ensure that its context, meaning and use can be properly determined. As Nathan (2010b: 196) states: '[M]etadata is the additional information about data that enables the management, identification, retrieval and understanding of that data. The metadata should explain not only the provenance of the data (e.g. names and details of people recorded), but also the methods used in collecting and representing it.' Notice that metadata is required not only for archiving but also for the very management, identification, retrieval, and understanding of the data within the documentation project once processing and value-adding is to be done. The way files are named and structured in folders is itself a type of metadata (see Nathan 2010b), and as Nathan and Austin (2004) argue, any knowledge added to the recordings (including transcription, translation, annotation, summary, index etc.) should be seen as 'thick metadata' (contrasted with the 'thin' cataloguing metadata often promoted in discussions of language documentation, e.g. by the E-MELD 'School of Best Practices in Language Documentation').

Nathan (2010b: 196) also proposes that: '[A]nother way to think of metadata is as meta-documentation, the documentation of your data itself, and the conditions (linguistic, social, physical, technical, historical, biographical) under which it was produced. Such meta-documentation should be as rich and appropriate as the documentary materials themselves.' It is my contention that alongside language documentation we need to develop a theory (and related practices) of language meta-documentation (or Meta-documentary Linguistics), the focus of which would be (to adapt the definition of Himmelmann 2006) the methods, tools, and theoretical underpinnings for setting up, carrying out, and concluding a documentary linguistics research project. It would be the documentation of the documentation research itself.

Some work on particular issues that are relevant here has been published in the language documentation literature, especially in relation to research ethics (Grinevald 2003, Dwyer 2006, Rice 2006, Thieberger and Musgrave 2006, Macri 2010), reciprocity and exchange (Yamada 2007, Czaykowska-Higgins 2009, Glenn 2009, Guerin and Lacrampe 2010, Leonard and Haynes 2010, Crippen and Robinson 2013), and researcher and community motivations (Dobrin 2008), which are part of what Dobrin and Berson (2011: 189) call 'the social processes set in motion by . . . research, from the conceptualization of fieldwork to the dissemination of its products'; however, no wider approach

[1] The documentary linguistics literature pays little attention to the role of still images in corpus creation. However, evidence from materials archived by documenters, e.g. at ELAR, suggests that they take large numbers of photographs and scans for a range of purposes (including documenting their field sites, recording setup, ceremonies, objects, and consultants and other people, and for copying fieldnotes and other documents, etc.).

or theorization has been undertaken. There are several reasons it would be valuable to do so:

- to develop good ways of presenting and using language documentations (what Woodbury 2011b calls 'making language documentations people can read, use, understand and admire')
- for future preservation of the outcomes of current documentation projects
- to assist with sustainability of the field of language documentation in terms of ensuring continuity of projects, people, and products
- helping future researchers learn from the successes and failed experiments of those currently grappling with issues in language documentation (see Gawne 2012 and comments by James Crippen)
- to document intellectual property contributions to projects, including those of community members, researchers, and others, along with their career trajectories, especially for more junior researchers (Conathan 2011a).

There are at least three possible directions that could be explored to strive towards a theory of meta-documentation:

1. *deductive approaches*: the postulation of axioms and theorems
2. *inductive approaches*: examination of current and past documentations (so-called 'legacy materials') to analyse practices and identify operating principles (as well as lacunae)
3. *comparative approaches*: examination of what other relevant and related fields have done in their meta-documentation to see what is applicable, and what not, to language documentation.

We discuss each of these in turn.

3.1 Deductive approaches

Since its establishment in the late 1990s language documentation has been dominated by declarative deductive approaches to recommendations for creating metadata which have been primarily influenced by library concepts (e.g. Dublin Core). Key metadata notions have been interoperability, standardization, discovery, and access (OLAC,[2] E-MELD, Good 2002, Farrar and Lewis 2007). However, the wider goals of language documentation (including the wider social goals relating to speaker community involvement) mean this is not powerful enough and we need, as P. Austin (2010a: 29) argues, to 'extend the concept of meta-documentation to include as full as possible documentation of the documentation project itself'. It appears that meta-documentation of at least the following aspects should be covered:

- the identity of the stakeholders and their roles in the project beyond the, so far, restricted concern to document people and roles such as 'speaker' and

[2] See the Appendix to this chapter for a listing of OLAC metadata terms and their definitions.

'recorder' (for a fuller but still incomplete listing see the OLAC roles given in Conathan 2011b: 245). For many projects other people, organizations, and institutions play a crucial role, e.g. funders, gate-keepers, validators of the research etc., but they and their roles tend to be neglected in metadata creation.

- the attitudes of language consultants, both towards their languages and towards the documentation project. These can of course change and develop over time and have a vital impact on the success or failure of a project, as well as the nature of the materials which can be collected and disseminated.
- the methodology of the researcher, including research methods and tools (see Lüpke 2010), and any theoretical assumptions encoded through abbreviations or glosses, as well as relationships with the consultants and the community (Good 2010 mentions what he called 'the 4 Cs': 'contact, consent, compensation, culture')[3]
- the biography and history of the project,[4] including the background knowledge and experience of the researcher and the main consultants[5] (e.g. how much fieldwork the researcher had done at the beginning of the project and under what conditions, what training the researcher and consultants had received), and how the project emerged and developed. For a funded project, the project biography would include the original grant application and any amendments, reports to the funder, e-mail communications with the funder, and/or any discussions with an archive, such as the reviews of sample data mentioned by Nathan (2010b). Both successful and unsuccessful aspects of the project biography should be included.
- any agreements entered into, whether formal or informal (such as a Memorandum of Understanding, payment arrangements, and any promises and expectations issued to stakeholders).

This kind of information is invaluable not only for the researcher and others involved in a project but also for any other future parties wishing to make sense of the project and its history and context. Unfortunately, linguists have typically been poor at recording and encoding this kind of information, meaning that work is often difficult with so-called 'legacy data', especially materials that only become available once the researcher has died (see Subsection 3.2 and Bowern 2003, P. Austin 2010b, Innes 2010, O'Meara and Good 2010). This is an area for further development and experimentation within language documentation theory and practice.

[3] This seems to correspond to Woodbury's (2011a, b) 'corpus theorization'.
[4] Note that OLAC (see the Appendix to this chapter) allows a date specification in the metadata for individual resources but is vague about the significance of such dates, defining it merely as 'a date associated with an event in the life cycle of the resource'.
[5] Conathan (2011b: 248) mentions biographical information about project participants but not the historical biography of the research project itself.

3.2 *Inductive approaches*

An inductive approach to meta-documentation would involve exploring current and past practices of language documenters to see what types of metadata they collect and notate within their projects. Here we report on two examples of such an approach: (1) a review of metadata practices in the Endangered Languages Archive (ELAR) at the School of Oriental and African Studies (SOAS) carried out by Nathan (2011), and (2) the main points from P. Austin (2010b), which look at the challenges of working with Australian Aboriginal legacy materials.

Nathan (2011) is a survey of metadata practices in forty-nine deposits in ELAR. He found that about 80 per cent of the most frequently occurring categories can be mapped to OLAC labels (see the Appendix to this chapter). However, depositors added richer specifications of other kinds of metadata information, including such things as parents' and spouse's mother tongues, speaker education levels, workflow status of materials, and terms in the researched languages (such as song titles or place names), or in other locally significant languages. Across the deposits examined some of these terms appeared frequently (e.g. 1 occurred in 20 of the 49 deposits); however, there were 613 terms which were unique and only occurred once in all the deposit descriptions, giving a 'long-tail' distribution (Anderson 2006). Nathan (2011) concludes that 'for endangered languages documentation, the metadata framework is to be *discovered*, *not predefined*, and the principle of the Long Tail is the opposite of focusing on the top 10–20 keywords . . . if supported and encouraged, documenters do produce diverse and more comprehensive metadata' (my emphasis). Nathan's review is suggestive of documenter practices for one archive, but needs further elaboration if it is to serve as a counterpoint to the deductive approaches which have dominated the field so far and which have emphasized standardization and metadata templates.

A second example of induction comes from P. Austin (2010b), which looks at issues arising from working with legacy materials on the Guwamu language from southern Queensland, Australia, collected in 1955 by the late Stephen Wurm. There are practical, technical, ethical, and political issues that this legacy data raises because of a lack of meta-documentation. Exploring these gives insights into what current documenters might wish to take into account for future users.

The Guwamu materials consist of: (1) fieldnotes of language elicitation (translations from English to Guwamu) collected from Willy Willis at Goodooga and comprising forty double-sided pages of notes with phonetic transcription and glosses in Hungarian shorthand, and (2) a short tape recording. At my request, the glosses were decoded and translated into English by Wurm and recorded onto tape in 1977. I copied the fieldnotes and added the English glosses (by transcribing Wurm's tape recording), resulting in a

138-page manuscript, a copy of which was deposited with the Library of the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS).[6] The following is a sample of data from the notes:

jama inda goammu ŋalgaŋanda?	Do you speak Guwamu?
baðarinj ŋalla	He is sick.
balgaru ŋunan ugwɛ:ilɛja	A few days ago I camped there.
balunj ŋadju ilu iðamanjgija juraŋu-nda	I will leave my axe here with you all.

Very little metadata was recorded with these materials, with the result that there are difficulties with them of several types:

1. Problems with the form of the original:
   a. The handwriting in the notes is sometimes difficult to decipher.
   b. Orthography – Wurm's transcription is not documented anywhere in the notes but appears to be similar to the International Phonetic Alphabet. It is quite surface phonetic but appears both to overdifferentiate (e.g. by recording gemination for consonants) and underdifferentiate (e.g. by failing to distinguish apico-alveolar and lamino-dental nasals).
   c. Word boundaries are sometimes incorrectly represented.
   d. There is sometimes cryptic glossing, or apparently wrong glossing.[7]
   e. Changing understandings over time of the language being recorded – Wurm was clearly working out the structure of Guwamu as he went along (and there are some comments in the fieldnotes which indicate his guesses about particular morphemes), so his transcription (and translation) varies from the first page to the last.[8]
2. Problems with the lack of context – we know nothing of how the material was recorded, what sessions took place, the background of the speaker and his involvement in and attitudes towards the project (on tape he sounds enthusiastic, at least when singing). No information is available about agreements entered into or any compensation or dissemination arrangements.
3. Problems of unclarity about protocol, i.e. access and usage rights to the materials in their various forms. The copy of Austin's notes at AIATSIS have the following access restrictions applied to them: 'Closed access – Principal's permission. Closed copying & quotation Principal's permission. Not for

[6] See www.aiatsis.gov.au/library/docs/langbibs/Guwamu_Kooma_July07.pdf.

[7] P. Austin and Crowley (1995: 60) give examples from work on legacy materials of such errors arising because the collector could not understand the consultants' accent or pronunciation, or because the semantics were misunderstood; we find instances of the latter in Wurm's notes but not the former.

[8] Bowern 2003 mentions that Gerhardt Laves began to analyse Bardi material collected in the 1930s as he was writing it down and made mistakes as a result, i.e. he did not write what he actually heard but what he thought he had heard. Also, Steele (2005: 84) comments on William Dawes' notebooks on the Sydney language: 'In order to be in a position to make some assessment of the soundness of an interpretation of a word, expression or sentence provided by Dawes, it is useful to have an idea of at which stage of his language learning an entry was created.'