elpublishing

Language Documentation and Description

ISSN 1740-6234

This article appears in: *Language Documentation and Description, vol 1*. Editor: Peter K. Austin

Introduction (LDD 1)

PETER K. AUSTIN

Cite this article: Peter K. Austin (2003). Introduction (LDD 1). In Peter K. Austin (ed.) *Language Documentation and Description, vol 1*. London: SOAS. pp. 6-14

Link to this article: http://www.elpublishing.org/PID/003

This electronic version first published: July 2014

Introduction

Peter K. Austin[1]

In this paper I wish to point to some issues that I think should be addressed to help us think about the approaches to be taken by field linguists in their endangered languages research.

1. What is language documentation and how does it differ from language description?

Linguists more or less understand what we mean by 'language description' (especially in the context of small or endangered languages) and the genres of standard description products are reasonably well conventionalised: dictionaries, grammars, text collections, and journal articles. There are publication outlets and evaluation and review processes for these, and descriptive linguists understand how their careers may be enhanced by adopting these genres. Other more experimental production has recently appeared in the context of emerging possibilities of multimedia (see Csató and Nathan below): web sites, CD-ROMs, talking and picture dictionaries etc. There are so far no standard genres for this kind of writing and no evaluation metrics ("how many pages is your web site if we print it out?" a colleague at my university was recently asked when applying for promotion and including his multimedia research in his publications).

Language documentation as a research area and an activity is not yet understood — what do we mean by 'documenting a language' as distinct from collecting data for linguistic description? Himmelmann (1998: 166) provides one view of this:

> "The aim of a language documentation is to provide a comprehensive record of the linguistic practices characteristic of a given speech community. Linguistic practices and traditions are manifest in two ways: (1) the observable linguistic behavior, manifest in everyday interaction between members of the speech community, and (2) the native speakers' metalinguistic knowledge, manifest in their ability to provide interpretations and systematizations for linguistic units and events. This definition of the aim of a language documentation differs fundamentally from the aim of language descriptions: a language description aims at the record of A LANGUAGE, with "language" being understood as a system of abstract elements, constructions, and

[1] My thinking on some of the issues discussed here has been influenced by discussions at various times with Colette Grinevald, David Nash, David Nathan, Nick Thieberger, David Wilkins, Peter Wittenburg, and Tony Woodbury. None of them is responsible for the content herein.

> rules that constitute the invariant underlying structure of the utterances observable in a speech community."[2]

Other views about documentation versus description were raised during a workshop organised by David Wilkins at the Max Planck Institute for Psycholinguistics, Nijmegen in 1995 entitled 'The Best Record', but none of these alternative views have been published. There is a gap in our intellectualising and theorising of this area of research and practice (cf. theorising (and rediscovery) of hypertext in the wake of the emergence of the World Wide Web). Tony Woodbury's contribution to this volume presents an overview of many of the relevant issues, including concrete examples from documentation projects that he has been involved in for a number of years.

It is also clear from recent experience that documenting a language with the aim of producing, eg. high quality annotated video on CD-ROM involves an understanding of performance and staging that is quite different from linguists' usual field experience of pointing a video camera at a 'language event' and pressing the on-button. So far, the fieldwork model in linguistics has been recording data for evidence (typically recorded 'raw' in the field and accompanied by a host of extraneous environmental noises) that can then be used to support the linguistic analysis and description. For multimedia documentation we require data for publication, with corresponding requirements about lighting, camera angles, and sound and image quality. Ideas on how to achieve this in the kinds of field situations field linguists work in need to be explored.

Also of importance is developing ways to collect fieldworkers' experience in documentation projects, theorise it and also draw from the wealth of knowledge and experience in, eg. computer-aided language learning (mostly however based on projects dealing with 'big languages'), to develop language documentation as a professional field. We also need to look closely at how theory and practice in information and communication technologies assist with this goal.

2. How do we evaluate language documentation?

We have conventional means of evaluating research in standard print format for academic or practical purposes — book reviews, citation indices, useability of dictionaries or school book materials, etc and conventional means for evaluating research in progress — conference presentations, poster sessions, workshops, seminars. What about language documentation, especially products in multimedia format? What evaluation metrics might we develop and how would these be implemented to help guide researchers as they work?

---

[2] Notice that this is not to be confused with the Saussurean distinction between *langue* and *parole* since the object of documentation is **knowledge** not simply behaviour.

3. Who are the key players in language documentation and what do they want?

An important issue for language documentation projects is: can we identify the **interested parties** and define their needs, expectations, roles, responsibilities, rights and relationships? For the HRELP Endangered Languages Documentation Programme, we have identified the following, each embedded within a legal and institutional framework (note that a particular individual may play one or more of these roles simultaneously or sequentially – see Wittenburg, this volume, regarding the DoBeS parties):

- Funding Agency
- Archiver
- Research Teams (including field linguists, anthropologists, ethno-botanists, IT specialists etc)
- Speaker Community
- Users
- General Public

Having established the parties and their role we may ask: what are the **key principles** by which the behaviour of each party is governed? Can we formulate guidelines on, eg. ethics, moral rights, relations with communities, archivist/researcher responsibility to future generations etc? Equally important is how these can be enforced, given that researchers in the past have tended to operate independently and without any professional requirements. Several bodies such as the Australian Linguistic Society and the American Anthropological Association have prepared such guidelines, however mostly they date from periods before the global publication and distribution regime of the world wide web.

A further issue to clarify is: what are the **key products** that each party wants? The kinds of output from research projects could be: a pile of data files, metadata, structured data sets in some standard format, web sites, print publications, CD-ROMs, software tools, or some combination of these. Often the kinds of products that communities want (such as practical dictionaries or school lessons) will be quite different from those desired by the academic research community, or the general public, or the funding agency. When there is conflict between the wishes of these groups, how are the **competing desires** of the parties to be resolved? (eg. what does the fieldworker do if they are expected to produce structured XML data by colleagues preparing metadata search tools, also expected to make a nice talking picture dictionary by the speakers?) What compromises are needed and what can we live with?[3] What information and communication technology tools can be developed to

[3] As David Nathan (2002) in a paper entitled "Don't mind the Grannies, feel the bandwidth" has recently pointed out, at the same time as flexible tools are being developed to increase the complexity of linguistic description ("the linguistic bandwidth") less attention is being paid to the desires of speakers for whom the language in use is paramount.

take compromise solutions and convert them to other formats (eg. an RTF to XML parser for a linguist who wants to maintain a dictionary in MS Word but make it more generally available)?[4]

4. What can we learn from the past and current practice about documenting languages and how do we plan strategically for the future?

Many of the corpora that exist for linguistic research are accidental, eg. extinct language inscriptions that just happen to survive, or fieldnotes from a one-day session with a speaker of a now extinct language. Current experience with language and cultural revival projects in Australia and North America can teach us a lot about the kinds of 'gaps' that exist in such corpora (eg. no information on phatic communion of greetings and leave takings, nothing about turn-taking in conversation or fillers, or how silence is used). It will be important to learn from and build upon the experience of field archives like the California Language Survey that have been in operation for 50 years[5].

An interesting question to consider is: if we know that there is now just one opportunity to document a language, what do we minimally collect and what tools and advice can we provide to field researchers to help them make the best corpus of data and metadata (given that we can not fully predict the future uses to which it might be put)? While there may not be a single one-size-fits-all recommendation that can be provided, there may be valuable suggestions for methods and guidelines that one can develop from past and current practices.

5. What are the professional consequences of doing language documentation?

The emerging field of language documentation is in need of professionalisation so that a new generation of scholars can be created. There will be training needs, eg. in information and communications technology, use of video, archiving practices, writing metadata etc. that will require the development of training programmes to distribute and transfer skills between participants. Linguists have traditionally acted as 'lone wolves' working alone, and sometimes in competition over the same language — for the future it will be important to encourage a team-oriented approach to research (including fully sharing 'control' with

[4] There are many examples of successful tools of this type, including David Nathan's eMu software that converts between markup formats, Chris Manning's Kirrkirr project that converted SIL backslash code Warlpiri dictionary files into XML, and the econv conversion tools developed at MPI Nijmegen (see www.mpi.nl/DOBES).

[5] I wrote to Leanne Hinton of the California Survey in mid-May 2003 asking for information about their experience but have not received a reply.

other researchers and speaker communities). Developments like the Electronic Cultural Atlas Initiative can potentially contribute to our understanding about collaborative models of research practice. There are also (software) tools that could be developed to assist with encouraging collaborative research, such as web-based front ends for multi-user databases, and we need experiments and theoretical approaches to these.

We also do not yet appreciate what the consequences are of undertaking this work for younger scholars. The ways that PhD programmes have been developed, especially in English speaking countries, mean that it will be difficult to meet the goals of language documentation as they are currently formulated while at the same time writing an acceptable dissertation, and doing so within the increasingly strict time limits that academic structures demand. A group of Australian PhD students raised this issue in an open letter to the Australian Linguistic Society membership in the May 2002 ALS Newsletter – see the Appendix below – however neither this society nor others that I am aware of have addressed this issue seriously. Given the amount of time language documentation demands in terms of transcription, annotation, and creation of metadata, it is important to think about how language documenters can advance their careers, or at least not set them back.

There is an important role for professional societies in responding to these new developments. There may well be opportunities for professional development, show-casing with constructive criticism, and publication, through the existing infrastructure of the professional societies (eg. LSA summer schools, conferences etc.)

The establishment of international networks and alliances to support the work of language documentation will also be a challenge for the future. Given that there are significant funds now coming on line to support this research effort from sources such as the Libet Rausing Charitable Fund, the Volkswagen Foundation, and the National Science Foundation, among others, we need to ensure that information and experience is shared and time and effort is not wasted rediscovering each other's errors. Recently there have been discussions about the formation of an international network of digital language archivists, and we can expect similar groups to emerge among language documenters in future years.

6. The papers

The papers collected here cover a wide range of topics within the general area of language documentation, especially from the perspective of endangered languages, however several themes are evident and reappear in a number of the papers, namely:

- what language documentation is and how it differs from linguistic description
- the role of speakers in documentation projects

- the role of new information and communications technologies, especially in the production of rich documentations using multimedia and computer-generated dictionaries
- what we can learn from history about language endangerment and shift, especially through detailed case studies
- the importance of training, both for the linguistics researcher communities and for the endangered languages communities themselves
- the importance of promoting understanding about endangered languages and language documentation, especially among the general public, and the ways that this can profitably done

Lisbet Rausing's paper was given at the launch of the Hans Rausing Project and overviews the main concerns of language documentation. David Crystal then addresses the question of what should be done now that linguists understand the current world ecology and the effects of language endangerment reasonably well. He replies that we must engage the interest of the general public, much in the way that the ecology movement aroused public consciousness about dangers to the physical environment and biodiversity, and that it should be done via the arts, especially theatre. Anthony Woodbury looks at the general topic of documentation and identifies it as an emerging area that brings together corpus studies, endangered languages concerns, neo-Whorfian interest in language diversity, and information technology. He argues that there are no single solutions to what is appropriate documentation and presents two rather different case studies to support his point. Colette Grinevald focuses on the speaker dimension of endangered languages work and outlines a typology of speakers and situations, also pointing to diversity of individual contexts.

The contribution by Eva Csató and David Nathan also focuses on speakers and communities, within the context of multimedia production. They present a case study from work they have done with the Karaim community of Lithuania and the resulting CD-ROM, which accompanies this volume, showing how speakers and their role must be included in project design at an early stage. William Foley explores the issue of literacy and spoken versus written language, showing that the traditional dichotomy between the two is complicated in small language communities, and that in fact certain characteristics of written language can appear in spoken texts that are prompted by a picture narrative. Foley's example is a reminder of the importance of genre and the need to pay attention to context and use in documentation.

Johanna Nicholas and Ronald Sprouse give a detailed description of a computer-aided multilingual dictionary project that they have worked on involving Caucasian languages. Their paper is a model of the ways in which properly applied information technology can truly assist both the linguist and the dictionary users. Peter Wittenburg outlines the framework of the DoBeS project on endangered languages documentation

sponsored by the Volkswagen Stiftung. This project has been running since 2000 and has already achieved impressive results in terms of defining documentation and archiving strategies and developing new computer tools to assist fieldworkers in team-based research groups.

The next papers take a smaller scale regional focus. Dan Everett overviews the linguistic diversity of the Amazon region and the research work that has been done there, both by himself and other local and international colleagues. E. Annamalai presents an overview of the situation in India, especially the role of the Central Institute of Indian Languages in the enormous task of documenting the daunting range of languages on the Indian sub-continent. Finally, Nicholas Ostler presents examples from history, namely Portuguese and English, to show that the fate of languages is not sealed and that languages that are today dominant were once themselves endangered. He notes however that contemporary language endangerement is taking place is a rather different social and economic environment than has ever been seen in the past, and that we must be sanguine about the future of the 50% or so of the world's languages that are in current danger. He concludes by outlining some of the ways that communities, governments and scholars can buttress languages to face an uncertain future.

Appendix — Open letter from postgraduates, published in May 2002 Australian Linguistic Society Newsletter

Given:

1. the recognition by linguists of the need for urgent work to record small ('endangered') languages, and
2. the availability of funding for that work, and
3. rapid advances in technological aids to doing that work,

then there is a clear need for the members of the Australian linguistic community to consider the following:

- The current PhD in linguistics makes no provision for language documentation beyond an academic grammar (in fact it positively discourages it).
- Documentation of a language with few speakers or with little prospect of being spoken in the next generation should be considered a suitable PhD topic in linguistics.
- This language documentation would produce information in the language in a form that makes it accessible to speakers and their descendants, and to the linguistic community. Typically this would include a grammar sketch, dictionary and texts.
- The form of the documentation would include as much information as possible, but would minimally provide audio and video recording of performance in the language. It could also include ethnobiological information such as pictures of plants and animals, their uses and names and so on.
- The document would include grammatical information, but not of the detail currently expected of a PhD in linguistics.
- The document would be produced using current standard tools (e.g. digital recording, text/audio linkage).
- The document would be presented in archive quality and placed in an appropriate repository to ensure its accessibility and usability into the future.
- This all entails training students in documentary techniques and linguistic data management.

In the long run it is the documentation that will prove more valuable for linguistic analysis than the traditional PhD. At present we have to rely on the writer of a PhD nearly 100% for some languages - and certainly 100% if the language is now gone.

While the current system values language analysis, it places no value on linguistic data management, nor on safely archiving recorded materials.

Louise Baird - ANU
Claire Bowern - Harvard
Nicolette Bramley - UC/ANU
Pascale Jacq - ANU
Anthony Jukes - Melbourne Uni
Doug Marmion - ANU
Stephen Morey - Monash Uni
Adam Paliwala - Sydney Uni
Carol Priestley - ANU
Adam Saulwick - Melbourne Uni
St John Skilton - Sydney Uni
Nick Thieberger - Melbourne Uni
Myfany Turpin - Sydney Uni