Language Documentation in the 21st Century

Prof Peter K. Austin
Endangered Languages Academic Programme
Department of Linguistics, SOAS
Outline

• What is language documentation?
• Establishing principles for the field
• Some current challenges
• Conclusions
Language documentation

• “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” (Himmelmann 1998)

• has developed over the last decade in large part in response to the urgent need to make an enduring record of the world’s many endangered languages and to support speakers of these languages in their desire to maintain them, fuelled also by developments in information and communication technologies

• essentially concerned with roles of language speakers and their rights and needs
Books and journals

- Gippert et al 2006 *Essentials of Language Documentation*. Mouton
- Tsunoda 2006 *Language endangerment and language revitalization: an introduction*
- *Language Documentation and Description* – 11 issues (2,000+ copies sold), 1 in prep
- *Language Documentation and Conservation* – 6 issues (on-line only)
- *Cambridge Handbook of Endangered Languages* 2011
- *Routledge Essential Readings* 2011
- *Oxford Bibliography Online* 2012
Main features (Himmelmann 2006:15)

- **Primary data** – collection and analysis of an array of primary language data to be made available for a wide range of users;

- **Accountability** – access to primary data and representations of it makes evaluation of linguistic analyses possible and expected;

- **Long-term storage and preservation of primary data** – includes a focus on archiving in order to ensure that documentary materials are made available to potential users now and into the distant future;
Main features (cont.)

- *Interdisciplinary teams* – documentation requires input and expertise from a range of disciplines and is not restricted to mainstream (“core”) linguistics alone

- *Cooperation with and direct involvement of the speech community* – active and collaborative work with community members both as producers of language materials and as co-researchers
A 21st century model

Woodbury 2011 enlarges concept of language documentation:
“creation, annotation, preservation and dissemination of transparent records of a language.”

and identifies several gaps in a Himmelmann-type approach:

“While simple in concept, it is complex and multifaceted in practice because:

• its object, *language*, encompasses conscious and unconscious knowledge, ideation and cognitive ability, as well as overt social behaviour;

• *records* of these things must draw on concepts and techniques from linguistics, ethnography, psychology, computer science, recording arts and more;
A 21st century model

- the *creation, annotation, preservation* and *dissemination* of such records pose new challenges in all the above fields, as well as information and archival sciences and;

- above all, humans experience their own and other people’s languages viscerally and have differing stakes, purposes, goals and aspirations for language records and language documentation”

He emphasises:

- *Diversity* of goals, purposes and outcomes
- Need for a *theory* of the documentary corpus
- Need for *accounts* of individual project designs
Challenges in 21st century

- Corpus taming
- Corpus theorisation and expression
- Archivism
- Standardisation stupidity
- Hokey-pokey hokum
- Meta-documentation
- Applied documentation
Challenge 1: Corpus taming

• Language documenters are often not well trained how to create and manage a corpus
• Without proper planning and workflow management a corpus can become uncontrolled and uncontrollable
• Begins with file naming conventions and organisation, metadata management, transfer protocols and good overall information management
Result – data dumps
Challenge 2: Corpus theorisation

*Corpus theorisation*: “the ideas according to which a corpus is said to cohere or ‘add up’”

Projects may aim to be comprehensive and diverse (and hence have to sample in some way) but they must take complementarity into account and think about gaps in existing corpora as well as community goals:

“corpus theorization, and indeed the very design and conduct of documentation projects, is also driven by social, aesthetic and humanistic values in the speech community itself, as well as those that develop within the emergent communities of practice conducting the documentation”
Challenge 3: Archivism

- Archivism: the archive tail wagging the documenter dog
- There is an inclination to treat quantifiable properties such as recording hours, data volume, and file parameters, and technical desiderata like ‘archival quality’ and ‘portability’ as primary criteria for assessing the aims and outcomes of language documentation
- Question to ELAR: “how many hours of video am I allowed to collect and archive for my ELDP project?”
New modes of expression

“Language documentation is going to need new genres and modes of expression to deal with its multimodality, hypertextuality, relationships between resources, and diversity of users and usages. These genres and modes of expression cannot be described or exhausted by the considerations of archiving-as-preservation … it seems that in general many documenters are struggling with formal aspects of their documentary work because of a late recognition by leaders in documentary linguistics that a good language documentation might be very much more than a set of dozens, hundreds, or thousands of files in archiveable formats.” (Nathan 2012)
Challenge 4: Recognising diversity

- Every documentation project is different from every other one because goals, contexts, people, languages, cultures, societies, expectations are different
- So projects must be unique and different from each other – no recipes, no commodification (Dobrin, Austin, Nathan 2009)
- Standards exist (eg. ISO-639-3) but must be sensibly used – do not standardise for the sake of standardisation, because it is supposed to be “best practice”
Challenge 5: Hokey-pokey hokum

You put your data in
You get your data out
You put your data in …
• Language documentation insists on the need for archiving documentary data and analysis
• Some grant agencies (eg. ELDP, Volkswagen) *require* it as part of the grant award
• But EL archives have a poor record of access and accessibility to documentary materials – more attention to preservation than to interfaces for users and channels of communication between depositors and users
• ELAR at SOAS addresses this in several ways – see David Nathan’s presentation
Challenge 7: Meta-documentation
Metadata

- metadata is *data about data*
  - for *identification, management, retrieval* of data
  - provides the *context and understanding* of that data
- carries those understandings into the *future*, and to *others* (and hence is important for archiving and preservation)
- reflects *knowledge and practices* of data providers
Metadata

• recommendations for creating metadata for language documentation have been primarily influenced by library concepts (e.g., Dublin Core), and key metadata notions have been interoperability, standardisation, discovery, and access (OLAC, EMELD, Farrar & Langendoen 2003).

• the goals of language documentation mean this is not powerful enough and we need a theory of metadata and meta-documentation = documentation of language documentation models, processes and outcomes
Meta-documentation methods

- meta-documentation requires reflexivity by linguists concerning their own documentary models, processes and practices, but should also draw on experiences from neighbouring disciplines (such as social and cultural anthropology, archaeology, archiving and museum studies), and from considerations that surface in the interpretation of past documentations (legacy materials) – cf. Good 2010
Missing meta-documentation categories

• identity of **stakeholders** involved and their roles in the project
• **attitudes** of language consultants, both towards their languages and towards the documenter and documentation project
• **relationships** with consultants and community
• **goals** and **methodology** of researcher, including research methods and tools (see Lüpke 2010), corpus theorisation (Woodbury 2011), theoretical assumptions embedded in annotation (abbreviations, glosses), potential for revitalisation
- **biography** of the project, including background knowledge and experience of the researcher and main consultants (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)

- for funded projects, includes original grant application and any amendments, reports to the funder, email communications with the funder and/or any discussions with an archive (e.g. reviews of sample data)
• **agreements** entered into – formal or informal (eg. Memorandum of Understanding, future compensation arrangements), and any **promises** and **expectations** issued to stakeholders

• **relationships** between this project and any others, past or present or future
Challenge 7: Applied documentation

Should documentation contribute to sustaining endangered languages and the communities who want to maintain and develop them? Are there links between documentation and language maintenance and revitalisation?

- Dimmendaal (2004: 84): “revitalisation, in my view, should not be given high priority. When individuals decide to give up their mother tongue, they usually have good reasons for doing so”;
- Blench (2007: 153): “almost by definition it is hardly worthwhile to spend limited resources on languages whose speakers are deserting them”
- Emergence of examples of applied language documentation, eg. LDD 11, Wuqu’ Kawoq.
it seems that in general many documenters are struggling with formal aspects of their documentary work because of a late recognition by leaders in documentary linguistics that a good language documentation might be very much more than a set of dozens, hundreds, or thousands of files in archiveable formats." (Nathan 2012)
Conclusions

• we need to move beyond our current models and become more reflexive and analytical about our practices
• we need to bring more of the social aspect of human life into language documentation and linguistic research (where it has been missing for the past 20 years of renewed interest in endangered languages)
Thank you!