Defining Language Documentation

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

• What is language documentation?
• How does it differ from language description (and from linguistic theory)?
• 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
What documentary linguistics is not

• it's not about collecting stuff to preserve it without analysing it

• it's not = description + technology

• it's not necessarily about endangered languages *per se*

• it's not a fad
DoBeS projects
ELDP funding

Projects 2003-2007
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
back to Language Documentation
Main features (Himmelmann 2006:15)

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

• *Explicit concern for accountability* – access to primary data and representations of it makes evaluation of linguistic analyses possible and expected;

• *Concern for 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.)

• *Work in interdisciplinary teams* – documentation requires input and expertise from a range of disciplines and is not restricted to mainstream ("core") linguistics alone

• *Close 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
The documentation record

• core of a language documentation is often taken to be a corpus of audio and/or video materials with transcription, annotation, translation into a language of wider communication, and relevant metadata on context and use of the materials

• lexico-grammatical analysis (description) and theory construction is contingent on and emergent from the documentation corpus (Woodbury 2003, 2011)
An example – Stuart McGill

- 4 year PhD project at SOAS
- documentation of Cicipu (Niger-Congo, north-west Nigeria) in collaboration with native speaker researchers
- outcomes:
  - a corpus of texts (video, ELAN, Toolbox)
  - 2,000 item lexicon
  - archive (956 files, 50Gbytes)
  - overview grammar (134 pages)
  - analysis of agreement (158 pages)
  - website, cassette tapes, books, orthography proposal and workshop
Woodbury 2003, 2011 says a good corpus is:

- **diverse**; diverse in situations; in participants—people carrying various different social roles; in channels such as speech, writing, e-mail; in speech genres, including conversation, narrative, oratory, verbal art, formal and informal interaction, and so on; and perhaps in different dialects or varieties or codes, if the community in question is multilingual, and if the documentary focus is on the community rather than one particular language or code.

- **large**. The technology is there for it to be arbitrarily large.
• ongoing, distributed, and opportunistic. It continually grows. Many people contribute to its development; documenters take advantage of any opportunity to record, videotape, or otherwise document instances of language use. Documentation projects must be designed to put easily available, easy-to-use, well-diffused technologies in the hands of as many people as possible, and to train them to make high quality recordings. This is the opposite of the traditional model, where someone from outside the community controls documentation and the means for documentation.
• **transparent.** They should be properly annotated. They should be useable by scholars 500 years from now. Everything should be competently translated into a language of wider communication; and transcriptional practices should be elucidated with links to phonetic and phonological data. Mark-up schemes should be elucidated and tested for intersubjective reliability. It is never adequate to collect lots of text on audio- or videotape and consider the documentation accomplished.

• **preservable, and portable.** Proper metadata information should be given about each item of data, whether text, audio, video, or any other medium. Data should be archived. It should migrate easily to the new technologies that emerge every few years.
ethical. Documentation must be carried out ethically; data ownership be protected and documenters should work with and respect the agendas of those with whom they are involved, especially those producing or owning the data, or having a hereditary or ancestral stake in it.
Components of documentation

- **Recording** – of media and text (including metadata) in context
- **Transfer** – to data management environment
- **Adding value** – transcription, translation, annotation, notation and linking of metadata
- **Archiving** – creating archival objects, assigning access and usage rights
- **Mobilisation** – creation, publication and distribution of outputs
Documentation and description
Documentation and description

- **language documentation**: systematic recording, transcription, translation and analysis of the broadest possible variety of spoken (and written) language samples collected within their appropriate social and cultural context

- **language description**: grammar, dictionary, text collection, typically written for linguists

Documentation and description

- documentation projects must rely on application of theoretical and descriptive linguistic techniques, to ensure that they are usable (i.e. have accessible entry points via transcription, translation and annotation) as well as to ensure that they are comprehensive
- only through linguistic analysis can we discover that some crucial speech genre, lexical form, grammatical paradigm or sentence construction is missing or under-represented in the documentary record
Documentation and description

- without good analysis, recorded audio and video materials do not serve as data for any community of potential users. Similarly, linguistic description without documentary support risks being sterile, opaque and untestable (not to mention non-preservation for future generations and useless for language support)
Workflow

Description

something happened

applied knowledge, made decisions

something inscribed

NOT OF INTEREST

cleaned up, selected, analysed

representations, lists, summaries, analyses

presented, published

Documentation

something happened

applied knowledge, techniques

recapitulates

recording

made decisions, applied linguistic knowledge

representations, eg transcription, annotation

archived, mobilised

FOCUS OF INTEREST
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

- **defines** and **constrains** audiences and usages for the data
- **all** value-adding to recordings of events involves the creation of metadata – all annotations (transcriptions, translations, glosses, pos tagging, etc.) are metadata (Nathan and Austin 2004)
Challenges

- Corpus taming
- ILG blindness
- Archivism
- Standardisation stupidity
- Meta-documentation
- Sustainability
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: ILG blindness

• Many linguists are blinded by multi-tier interlinear glossing, believing it to be the “gold standard” of annotation

• Fallacy of “when all you have is a hammer everything looks like a nail” (Toolbox and ELAN as windows (blinkers) on world of documentation)

• Often better as first pass to do audition annotation – real time time-aligned summary of recordings, notating genre and content as searchable text (eg. using Transcriber, Excel)

• Selective ILG is better because of labour intensive nature – typically 100:1 for morpheme-by-morpheme glossing
Challenge 3: Archivism

- The archive tail wagging the documenter dog
- The 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 ELDLP project?”
Challenge 4: Standardisation stupidity

- 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 cookie cutter models, 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: Meta-documentation
Metadata revisited

- 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 (eg. 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, largely lacking until now
Meta-documentation goals

- developing good ways of **presenting** and **using** language documentations
- future **preservation** of the outcomes of current documentation projects
- **sustainability** of field
- helping future researchers **learn** from the successes and failed experiments of those presently grappling with issues in language documentation (Austin 2010)
- documenting **IP contributions** and career trajectories (Conathan 2011)
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 (Good 2010 mentions what he called ‘the 4 Cs’: ‘contact, consent, compensation, culture’);
- **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. the reviews of sample data)
• **agreements** entered into – formal or informal (e.g. 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 6: Sustainability

- we do not have good models for sustainable documentation projects
- we understand sustainability of archived data but how do we sustain projects and relationships beyond the typical 3-5 year academic life cycle?
- how can documentation contribute to sustaining endangered languages and the communities who want to maintain and develop them? What are the links between documentation and language maintenance and revitalisation? (This is sadly under-theorised.)
• we need to move beyond our current models and be more reflexive and develop better theorisation

• 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!