Defining language documentation

Peter K. Austin
Department of Linguistics
SOAS, University of London
Outline

- What is language documentation?
- Background and developments
- Documentation versus description
- Project design, workflows and skills
- Outcomes
What is 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 20 years 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, media and communication technologies

- concerned with roles of language speakers and communities and their rights and needs
What it 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 passing trend
Evidence

Student interest

- 140 students graduated from SOAS MA in Language Documentation and Description 2004-14 – currently 20 are enrolled
- 10 graduates in PhD in Field Linguistics – 20 currently enrolled
- other documentation programmes, eg. UTAustin have similar experience
- Summer schools (3L, InField/Collang), training courses

Publications

- Gippert et al 2006 Essentials of Language Documentation. Mouton
- Cambridge Handbook of Endangered Languages 2011
- Routledge Essential Readings 2011
- Oxford Bibliography Online 2012
- Language Documentation and Description – 12 issues (paper and online)
- Language Documentation and Conservation – 6 issues (online only)
Big money – DoBeS projects
Big archives – ELAR at SOAS
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

- Outcome is **annotated and translated corpus** of archived representative materials on a language
A 2010 example – Stuart McGill

- 4 year PhD project at SOAS, plus 2 year post-doc
- 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
McGill Cicipu corpus

- 68 yes
- 69 Hausa has swept them away
- 70 yes
- 71 if they want proper Cicipu, it's only the former people
- 72 m-hm, not the children of now
- 73 Cicipu only the former people
- 74 yes
- 75 yes
- 76 yes
- 77 one place, doing their work in one place
- 78 they would do it together

- óó
- ṭikóọ́ ṭihúdo ṭe
- yes
- Hausa has swept them away
- yes

- hūuw ɓi ɗanà ṭikóọ́ kóò?
- they speak it like Hausa?
## Cicipu annotations

<table>
<thead>
<tr>
<th>Texts.txt</th>
<th>Dictionary.txt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>svtnm001.099</td>
</tr>
<tr>
<td>Start Time</td>
<td>208.260</td>
</tr>
<tr>
<td>End Time</td>
<td>209.450</td>
</tr>
<tr>
<td>Speaker</td>
<td>Teni</td>
</tr>
<tr>
<td>Text</td>
<td>tikg5 ti- hudy : L -H - L re</td>
</tr>
<tr>
<td>Morphemes</td>
<td>k5g5 ti- hudy : L -H - L re</td>
</tr>
<tr>
<td>Glass</td>
<td>NC6- Islam AGH6- sweep -RLLsp 3PP.PRO nc- n agv- v -vtonep pro</td>
</tr>
<tr>
<td>Part of Speech</td>
<td><strong>Hausa person</strong></td>
</tr>
<tr>
<td>English Gloss</td>
<td><em>empty</em></td>
</tr>
<tr>
<td>Part of Speech</td>
<td><em>empty</em></td>
</tr>
<tr>
<td><em>empty</em></td>
<td>8/2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Texts.txt</th>
<th>Metadata.txt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>svtnm001.100</td>
</tr>
<tr>
<td>Start Time</td>
<td>210.260</td>
</tr>
<tr>
<td>End Time</td>
<td>210.960</td>
</tr>
<tr>
<td>Speaker</td>
<td>Teni</td>
</tr>
<tr>
<td>Text</td>
<td>oo</td>
</tr>
<tr>
<td>Morphemes</td>
<td>of</td>
</tr>
<tr>
<td>Glass</td>
<td>yes</td>
</tr>
<tr>
<td>Part of Speech</td>
<td>intj</td>
</tr>
<tr>
<td>English Gloss</td>
<td><em>empty</em></td>
</tr>
<tr>
<td>Part of Speech</td>
<td><em>empty</em></td>
</tr>
<tr>
<td><em>empty</em></td>
<td>7/8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metadata.txt</th>
<th>\audio_visual/svtm001.wav</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Discussion of chieftaincy</td>
</tr>
<tr>
<td>Contributor</td>
<td>GDM ; JN ; King</td>
</tr>
<tr>
<td>Audio File</td>
<td>\audio_visual/svtm001.wav</td>
</tr>
<tr>
<td>Recording Place</td>
<td>King's guest house, Korisino</td>
</tr>
<tr>
<td>Location</td>
<td>Mali's guest hut in Kadaada</td>
</tr>
<tr>
<td>Location</td>
<td>At the grave of Norman Biggs, Sakaba</td>
</tr>
<tr>
<td>Location</td>
<td>On the Ukula (Maburya) mountain</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metadata.txt</th>
<th>\audio_visual/svtm001.wav</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Greeting the Malian</td>
</tr>
<tr>
<td>Contributor</td>
<td>BB ; MK</td>
</tr>
<tr>
<td>Audio File</td>
<td>\audio_visual/svtm001.wav</td>
</tr>
<tr>
<td>Recording Place</td>
<td>At the grave of Norman Biggs, Sakaba</td>
</tr>
<tr>
<td>Location</td>
<td>At the grave of Norman Biggs, Sakaba</td>
</tr>
<tr>
<td>Location</td>
<td>On the Ukula (Maburya) mountain</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metadata.txt</th>
<th>\audio_visual/svtm001.wav</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Norman Biggs' grave</td>
</tr>
<tr>
<td>Contributor</td>
<td>MY</td>
</tr>
<tr>
<td>Audio File</td>
<td>\audio_visual/svtm001.wav</td>
</tr>
<tr>
<td>Recording Place</td>
<td>At the grave of Norman Biggs, Sakaba</td>
</tr>
<tr>
<td>Location</td>
<td>At the grave of Norman Biggs, Sakaba</td>
</tr>
<tr>
<td>Location</td>
<td>On the Ukula (Maburya) mountain</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metadata.txt</th>
<th>\audio_visual/svtm001.wav</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Ukula mountain</td>
</tr>
<tr>
<td>Contributor</td>
<td>MoMu ; SDT (Sani the</td>
</tr>
<tr>
<td>Audio File</td>
<td>\audio_visual/svtm001.wav</td>
</tr>
<tr>
<td>Recording Place</td>
<td>TMG's compound in Ka'ingawa KaGaladima</td>
</tr>
<tr>
<td>Location</td>
<td>TMG's compound in Ka'ingawa KaGaladima</td>
</tr>
<tr>
<td>Location</td>
<td>TMG's compound in Ka'ingawa KaGaladima</td>
</tr>
<tr>
<td>Location</td>
<td>TMG's compound in Ka'ingawa KaGaladima</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metadata.txt</th>
<th>\audio_visual/svtm001.wav</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>Interview about the old</td>
</tr>
<tr>
<td>Contributor</td>
<td>MoMa ; TMG ; Yaaki</td>
</tr>
<tr>
<td>Audio File</td>
<td>\audio_visual/svtm001.wav</td>
</tr>
<tr>
<td>Recording Place</td>
<td>TMG's compound in Ka'ingawa KaGaladima</td>
</tr>
<tr>
<td>Location</td>
<td>TMG's compound in Ka'ingawa KaGaladima</td>
</tr>
<tr>
<td>Location</td>
<td>TMG's compound in Ka'ingawa KaGaladima</td>
</tr>
<tr>
<td>Location</td>
<td>TMG's compound in Ka'ingawa KaGaladima</td>
</tr>
</tbody>
</table>
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

- **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.

- 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-preservable for future generations and useless for language support).
Workflow differences

**Description**

- Something happened
- Applied knowledge, made decisions
- *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
As a consequence

- Documentation needs an epistemology for media capture – audio and video recording (see Nathan)
- Need to pay attention for good practices in recording – eg. microphone choice and spatiality in audio, framing-lighting-editing for video (“recording arts”)
- Concern for socio-cultural context (“ethnography of speaking”)
- Concern for data structuring and data management – eg. ‘portability’, relational modeling, XML
- Concern for ‘standards’ and cross-project comparability, especially typology and data mining
- Concern for ethics of research – documentation collects language use in “intimate” personal contexts, impacts on potential users and uses of documented speech events
• Use of tools to support workflow components, eg. transcription, annotation, metadata management
• Descriptive linguists tend to be weak and/or untrained in these areas
• Woodbury 2011 notes that language documentation “must draw on concepts and techniques from linguistics, ethnography, psychology, computer science, recording arts and more” (where “more” includes history, archiving, museum studies, project management, creative writing, social media)
Woodbury (2003:46-47) – a good corpus

- **diverse** — containing samples of language use across a range of genres and socio-cultural contexts, including elicited data
- **large** — given the storage and manipulation capabilities of modern information and communications technology (ICT), a digital corpus can be extensive and incorporate both media and text
- **ongoing, distributed, and opportunistic** — data can be added to the corpus from whatever sources that are available and be expanded when new materials become available
• **transparent** — the corpus should be structured in such a way as to be usable by people other than the researcher(s) who compiled it, including future researchers

• **preservable, and portable** — prepared in a way that enables it to be archived for long-term preservation and not restricted to use in particular ICT environments

• **ethical** — collected and analysed with due attention to ethical principles and recording all relevant protocols for access and use.
Recording

- **audio** – basic and familiar in modern linguistic work. Important considerations: environment, equipment choice, microphones, monitoring, file type (wav not mp3 generally recommended)

- **video** – immediate, rich in authenticity, multi-dimensional in context, great interest to communities, can be produced independently by community members BUT more difficult to produce, process, access without time-aligned annotation, transfer, store and preserve

- **text** – compact, stable, easy to store, access and index, can express hypertextual links to other text and media BUT relies on literacy and is less rich than audio/video
• *metadata* – data about the data: needed to identify, manage, retrieve data. Provides context and understanding of data to oneself and others. Types:
  • Cataloguing — identifying and locating data, eg. language code, file id, recorder, speaker, place of recording, date of recording etc
  • Descriptive — kind of data found in a file, eg. abstract/summary of file contents, knowledge domain represented
  • Structural — specification of file organisation, eg. textfile is a bilingual dictionary
  • Technical — file format, kind of software needed to view, preservation data
  • Administrative — work log, intellectual property rights, moral rights, access and distribution restrictions
• *meta-documentation* – documentation of language documentation models, processes and outcomes, goals, methods and conditions (linguistic, social, physical, technical, historical, biographical) under which the data and analysis was produced (should be *as rich and appropriate* as the documentary materials themselves)
Adding value

- requires decision making (selection, editing, choice of method and theory) and is very time consuming (eg. annotation can be 100:1 in terms of time required)

- linguistic value adding (‘thick’ meta-data):
  - *transcription* – textual representation of audio signal (orthographic, phonemic, phonetic) typically time-aligned to media
  - *annotation* – overview, code, morphological, grammatical, semantic (‘gloss’), syntactic, pragmatic, discourse. Fixation among documenters on ‘interlinear glossing’, cf. overview annotation/summary
many documenters believe that interlinear glossing is the ‘gold standard’ of annotation but it is very time-consuming and illegible to non-linguists – overview annotations may be a preferred as a primary goal: ‘roadmap’ or index of a recording – approximately time-aligned information about what is in the recording, who is participating, and other interesting phenomena.
Item 408: Oral Literature Collection, Tape 343, Side B. Robert Zuboff (Kak’weidí clan, Kaakáakw Hít) and Susie James (Chookaneidi clan, T’akdeintaan yádi), July 27, 1972; interviewed by Nora Marks Dauenhauer, migrated from reel to CD. Length 60:14. **Content by DK:** story of how the Sea Otter came to be is told, 0-4:15; raven sounds are given by Zuboff, and their meaning/use, 4:16-11:10; Zuboff tells a story about a man who became an invisible man (tlékanáa) (13:24); 11:11-13:24; story of a man named Naawan that bit the tongue off of a raven, 13:25-16:09; general conversation and questions about Tlingit phrases, 16:10-19:57; story of a man named Gáneix, 19:58-21:40; discussion about language and storytelling, mention of the Salmon Boy story, 21:41-24:12; Zuboff tells the story about the Woman that Raised the Wood Worm, attributes the story’s people, 24:13-27:34; Susie and Nora talk, Susie speaks about the Man Who Commanded the Tides (Yookis´kookeik) and his sister and raven. She then tells the story of bringing in the house that was way out on the ocean and how raven got the octopus tentacle to bring in the house. She then talks about the type of resources that were in the house but not in detail. She mentions the whale, cod etc. She then goes back to the man who commanded the tide and rescues his mother by placing her in the skin of a black duck, 27:35 to the end of the recording. Notes on file.
Tools for value adding

- application programs, components, fonts, style sheets, and document type definitions (DTD).

- application programs:
  - *general purpose* software – user must design data structures and manipulation routines, eg. LibreOffice, MS Office (Word, Excel, Access)
  - *specific purpose* software – designed for particular tasks, eg. Transcriber, ELAN, Arbil, Flex, Toolbox

- Important: design and use a work flow that enables data transfer (export, import) without loss/corruption of encoded knowledge
Archiving

A digital language archive:

- is a trusted repository created and maintained by an institution with a commitment to the long-term preservation of archived material
- has policies and processes for acquiring, cataloguing, preserving, disseminating, and format/content migration
- is a platform for building and supporting relationships between data providers and data users
General archiving functions

- advise
- acquire
- preserve
- add value
- provide access
- provide collaboration channels
- develop trust
And *endangered* languages archiving?

- extremely diverse context – languages, cultures, communities, individuals, projects
- typical source - fieldworkers
- typical materials - documentation
- difficult for archive staff to manage
- sensitivities and restrictions
What can a language archive offer?

- **Security** - keep your electronic materials safe
- **Preservation** - store your materials for the long term
- **Discovery** - help others to find out about your materials, and you to find out about users
- **Protocols** - respect and implement sensitivities, restrictions
- **Sharing** - share results of your work, if appropriate
- **Acknowledgement** - create citable acknowledgement
- **Mobilisation** - create usable language materials
- **Quality and standards** - advice for assuring your materials are of the highest quality and robust standards
What do depositors have to do?

- select and contact an archive
- prepare materials
  - select
  - structure
  - suitable encodings and formats
  - metadata, metadocumentation, agreements
- send materials to archive(s)
- work with archive during curation etc.
- ongoing management, updating, dissemination
- ongoing collaboration with users
What can you archive?

- media - audio, video
- graphics - images, scans
- texts - fieldnotes, grammars, description, analysis
- structured data - aligned and annotated transcriptions, databases, lexica
- metadata, metadocumentation - contextual information about the materials, both structured and unstructured
Archive objects

- an ‘object’ could be a file, a set of files, a directory, or a set of files with their relationships explicitly defined
- archives use a set principle, at ELAR called “bundles” and at DoBeS called “sessions”
- a bundle or session contains one or more files (media, image(s), text) together with metadata about them and their relationships (eg. in a spreadsheet)
Example – deposit home page
Example – file deposit pages
Mobilisation

- Creation of usable outputs for a range of different audiences, e.g. multimedia websites, sub-titled video, apps
- There are tools to help with this (LexiquePro, CuPed) and people working on app development who can help, e.g. Ma! Iwaidja
Conclusions

- Language documentation: relatively new sub-field of linguistics aimed at producing well-structured archived annotated corpora and associated resources (metadata, meta-documentation, grammar/dictionary/text, mobilisation products, especially for community use)

- World-wide development with increasing numbers of people working in this framework and increasing interest, especially for endangered languages and language revitalisation

