Documentary and meta-documentary linguistics

Peter K. Austin

Endangered Languages Academic Programme
SOAS, University of London

ICLDC, Hawaii
7 February 2011

www.hrelp.org
Documentary linguistics

- Himmelmann (2006:v): the subfield of linguistics ‘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’

- Woodbury (2011:1): ‘the creation, annotation, preservation, and dissemination of transparent records of a language’

- metadata plays a crucial role
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)
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
Types of metadata

- creator’s / delegate’s details
- **descriptive** metadata – content of data
- **administrative** metadata – eg. date of last edit, relation to other data
- **preservation** metadata – character encoding, file format
- **access** and **usage** protocols – eg. URCS
- metadata for **individual** files or **bundles**

- Metadata can apply at various levels
DoBeS model

DoBeS
- Project
  - Corpus
    - Session
      - Item
      - Item
    - Session
      - Item
      - Item
  - Session
    - Item
    - Item
- Project
  - Corpus
    - Session
      - Item
      - Item
    - Session
      - Item
      - Item
  - Session
    - Item
    - Item
How do we store metadata?

- **in our heads** – problem: degrades rapidly and not preservable or portable
- **on paper** – problem: not easily searchable or extensible
- **within files** (headers) – problem: not easily searchable or extensible
- **in file/folder names** (eg. SasJBpka09-12_int03.wav) – problem: difficult to maintain, breaks easily, not all semantics can be expressed
- **in a metadata system**
Metadata systems

- free text
- structured text (e.g., Word tables, XML, Toolbox)
- spreadsheet (e.g., Excel)
- database (e.g., Filemaker Pro, Access, MySQL)
- metadata manager (IMDI, SayMore)

- or some combination of these that is usable, flexible and sufficiently expressive
Meta-documentation

• Nathan (2010): ‘[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.’ [emphasis added]

• meta-documentation = documentation of language documentation models, processes and outcomes
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
Approaching meta-documentation theory

- **deductive approach:** postulation of axioms and theorems
- **inductive approach:** examination of current and past documentations (so-called ‘legacy materials’) to analyse practices and identify operating principles (as well as lacunae)
- **comparative approach:** examine what other relevant and related fields have done in their meta-documentation, to see what is applicable and what not to documentary linguistics
Deductive: metadata formats so far

- **common or standard:**
  - IMDI (ISLE Metadata Initiative, DoBeS) – rich, for corpus management
  - OLAC (Open Language Archives Community) – compact, for retrieval
  - EAD (Encoded Archival Description), others

- individual organisations (eg. ELAR, AILLA, Paradisec) have developed their own sets and/or allow depositor’s own metadata

- but there is a yawning gap in coverage
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 (eg. 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 (eg. the reviews of sample data mentioned by Nathan 2010)

• **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
Inductive: current and past approaches

- Nathan 2011 survey of metadata practices by ELAR depositors

- Austin experiences of working with S. A. Wurm’s legacy materials on New South Wales languages

- Bowern experiences with Gerhard Laves materials on Bardi, Western Australia
Nathan 2011 overview

- collected information from about 50 deposits
- collected metadata categories, with illustrative data values
  - 37 deposits fully extracted
  - 7 partially extracted
  - 12 had no metadata
  - 5 were IMDI – extracted the key-value pairs only
About 80% of most frequently occurring categories can be mapped to OLAC.

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>language</td>
<td>Subject.language</td>
</tr>
<tr>
<td>date</td>
<td>Date</td>
</tr>
<tr>
<td>description</td>
<td>Description</td>
</tr>
<tr>
<td>id</td>
<td>Identifier</td>
</tr>
<tr>
<td>speaker</td>
<td>Contributor</td>
</tr>
<tr>
<td>title</td>
<td>Title</td>
</tr>
<tr>
<td>format</td>
<td>Format</td>
</tr>
<tr>
<td>type</td>
<td>Type</td>
</tr>
<tr>
<td>creator</td>
<td>Creator</td>
</tr>
<tr>
<td>file name</td>
<td>Identifier</td>
</tr>
<tr>
<td>notes</td>
<td></td>
</tr>
<tr>
<td>rights</td>
<td>Rights</td>
</tr>
<tr>
<td>duration</td>
<td>Coverage</td>
</tr>
<tr>
<td>content</td>
<td>Description</td>
</tr>
</tbody>
</table>
Other categories:

- detailed locations
- metadata in Spanish
- indigenous genres and titles (eg of songs)
- consultants’ parents’ and spouse’s mother tongues, birthplaces
- number of children, their language competence
- L2, L3 and competencies
- languages heard
- clan/moiety
- consultants’ occupation
- consultants’ education level
• date left home country
• photos (/captions) of consultants, field sessions etc
• equipment
• microphone
• workflow status
• naming and organisational codes and principles
• recorder/linguist experience level
• biography and project description ("meta-documentation")
<table>
<thead>
<tr>
<th>Term frequency</th>
<th>Number of terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>51</td>
</tr>
<tr>
<td>1</td>
<td>613</td>
</tr>
</tbody>
</table>
Conclusions of Nathan 2011

• “if supported and encouraged, documenters do produce diverse and more comprehensive metadata”

• “for endangered language documentation, the metadata framework [= theory of meta-documentation PKA] is to be discovered, not predefined”
A legacy example: Guwamu project

- Stephen Wurm’s fieldnotes of language elicitation (translations from English to Guwamu) collected from Willy Willis in Goodooga 1955; 40 double-sided pages of notes with phonetic transcription and glosses in Hungarian shorthand; short tape recording
- glosses decoded by Wurm and recorded on tape in 1977; fieldnotes copied and glosses added by Austin 1977, 138 pages, copy deposited at AIATSIS
jama inda goamu ŋalganda? Do you speak Guwamu?
bađarinj ŋalla He is sick.
balgaru ŋunan ugwε:ileja A few days ago I camped there.
balunj ŋadju ilu iđamanjgi jà juraŋu-nda I will leave my axe here with you all.
Meta-documentation issues: form

- orthography – Wurm’s transcription is not documented but appears to be similar to IPA – quite low level phonetic but both overdifferentiates (eg. recording gemination for consonants) and underdifferentiates (eg. failing to distinguish apico-alveolar and lamino-dental nasals)
- shorthand notations – Wurm’s glosses are mostly in Hungarian shorthand
- word boundaries sometimes incorrect
• sometimes cryptic glossing, or apparently wrong glossing
• changing understandings over time of the language being recorded – Wurm clearly was 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 varies from the first page to the last
Meta-documentation issues: context

- **stakeholder** issues: we know nothing of how the material was recorded, what sessions took place, the background of the speaker and his involvement in the project (on tape he sounds enthusiastic, at least when signing). No information is available about agreements entered into or any compensation arrangements.

- problems of unclarity about **protocol**, ie. access and usage rights to the materials in their various forms. The copy of Austin’s notes at AIATSIS have access restrictions: “Closed access - Principal's permission. Closed copying & quotation Principal's permission. Not for Inter-Library Loan”
Comparative: looking at other fields

- much work needs to be done here, but see Hanks 2011 and Good and Ember 2011 presentations at LSA in Pittsburgh for some beginnings
- archaeology (especially that influenced by Hodder 1999) has been more reflexive about its practices in the past 10 years than language documentation has (e.g. daily field diaries, debates about raw (field reports) vs. cooked (academic papers) etc.
Conclusions

- documentary linguistics has not paid sufficient attention so far to the nature, functions and expressive power of metadata. It has no theory of metadata.
- preoccupations with standardisation, driven by typologists and ‘data scrapers’ and fetishisation of interlinear glossing in particular have narrowed attention
- we need more reflexivity and exploration of meta-documentation in order for the field to develop further in the future
Thank you!