Anthony Aristar, Helen Aristar-Dry
LINGUIST List, Eastern Michigan University
Jeff Good
University at Buffalo, The State University of New York
• Few of us have the luxury of “just” documenting.

• We also need to mobilize our data.

• **Data mobilization**: The transfer of data from archival form to a form where it can be put to a concrete use.

• Most typically associated with community support, but also relevant to other contexts.
Mobilization for who?

- Many parties have a stake in language data
  - The data collector
  - The language community
  - The linguistics community
  - The wider academic community
  - The general public
• Two parties stand out
  • The data collector (here, “the linguist”)
  • The language community (here, “the community”)

• If the needs of each of these is not supported, the documentary efforts will fail

• Use of these heuristic categories glosses over massive variation among researchers and among communities
• Linguists need
  • Money for food and shelter
  • Money for research
• They must mobilize data in ways that will support these things
• In 2008, this more or less means mobilization for publication
What communities need

• Language communities need
  • Money for food and shelter $$
  • Money for community improvement
• Orthographies
• Pedagogical materials
• Dictionaries, texts
• Translations
• Academic support

June 25, 2008

Data Management & Archiving: Data types
Finding a balance

- It takes time to find the right balance between research and community interests.
- A one-size-fits-all approach is simply impossible.
- Thoughtful data management can help.
  - Some formats are more readily repurposed than others.
  - Multiple outputs can become relatively cheap.
• The beginning of the process is not technological, but personal

• Good communication must always be maintained between the linguist and the community

• Expectations and plans should be made explicit

• Ongoing discussion will refine those expectations and plans
The linguist should probably not develop an orthography.

But the linguist can use an electronic lexicon to help the community find a good set of words for testing an orthography.

And the linguist can include orthographic representations in their lexicon—or create automatic conversion scripts to an orthography.
• Outline planned contributions early
• Consult with experienced individuals about technological choices and availability of technology in the community
• Make community needs a factor in determining your implementation choices
• If your choices are not ideal from a community perspective, communicate this and plan a workaround
I’ve used Filemaker to make lexicons but it’s unsuitable as a format for the community. So, I should either choose a different tool the community can use or develop a plan to turn the Filemaker lexicon into an appropriate community format.
Conversions from research to community formats will almost always be possible

Pretty much any need you have will be technologically possible

The main problems are economic and social

Conversion costs money

Conversion takes (someone’s) time

Thus, planning ahead is necessary
• Possible community-oriented outputs
  • Dictionary
  • Orthography
  • Texts
  • Alphabet books/language primers
  • Interactive pedagogical multimedia
  • Bible translations
Output formats can take on many forms

- Print publications
- Interactive websites
- CD-ROMs

Considerations are functional and social

- An attractive book is prestigious
- A website is modern
Repurposing: Transformation of a resource from a format supporting one use to a format supporting another use

Digital formats allow for repurposing with an ease that was previously unimaginable

Markup formats coding logical structure, like XML, are particularly good for this
Linguist’s needs
- Easy searching of documented language
- Lots of grammatical information

Community needs
- Two-way searching
- Non-technical presentation
- Semantic fields
Good for linguist’s searching, not orthographic
Good for linguist and community
Will it facilitate reverse lookup?
My lexicon

Naki-Lexicon

nyam  nyam
L    ML
n
animal, meat

Guthrie CS  BLR Index
1909 a    3180

Notes

Entry ID
Naki73191661111

Browse
Some “notes” may be useful; others won’t be
<table>
<thead>
<tr>
<th>nyam</th>
<th>nyam</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>ML</td>
</tr>
</tbody>
</table>

animal, meat

Guthrie CS 1909 a
BLR Index 3180

Notes

Entry ID
Naki73191661111

Checked tone ☒ Yes ☐ No ☐ Other...
Transform with merge

Lexical Database
Transform with merge
Transform with merge

Lexical Database

Transform

Academic Dictionary
Transform with merge

- Comm. Database
- Lexical Database
- Transform
- Academic Dictionary
Transform with merge

Comm. Database → Merge and Transform → Lexical Database → Transform → Academic Dictionary
Transform with merge

Comm. Database → Merge and Transform → Dictionary / Thesaurus

Lexical Database → Transform → Academic Dictionary
A rich database

Rich Lexical Database

Comm. Transform

Acad. Transform
A rich database

Rich Lexical Database → Comm. Transform → Academic Dictionary

Rich Lexical Database → Acad. Transform → Academic Dictionary

Academic Dictionary / Thesaurus
• Less work in creation of initial lexical database, more work in creation of a community dictionary

• More work in creation of initial lexical database, less work in creation of community dictionary

• Data management techniques can’t tell you how to help the community

• But, they can make the job easier once a decision has been made
For unwritten languages, linguistic work can facilitate development of an orthography.

In particular, a phonemic analysis can be useful.

Orthographies may be non-ideal for analysis:
- They may not mark all relevant contrasts.
- They may make use of conventions hindering search and collation.
• Contrast ∫ and sh

• ∫ unambiguously codes [ʃ]

• sh may encode or [ʃ] or [sh]

• bishop vs. mishap

• This is not a problem for native speakers, but it can be for learners or linguists
• If a community is to use a linguist’s work, it must be able to use linguist’s transcriptions
• The simplest case: Linguistic transcription and community orthography are the same
• A simple case: Linguistic transcription can be straightforwardly mapped to orthography
• Hardest cast: Human intervention required for some parts of the mapping
Easy mappings

Linguist
a b c d e

Community
а б в Ъ е

Linguist
c k e

Community
а б к е
Easy mappings

Linguist

Community

Linguist

Community

June 25, 2008
Data Management & Archiving: Data types
Easy mappings

Linguist  Community  Linguist  Community

a  —>  a  
b  —>  б  
c  —>  Ц  
d  —>  А  
e  —>  е

a  —>  а  
b  —>  б  
c  —>  к  
k  —>  е
Harder mapping

Linguist

Community

a b c e

а б в с е
Linguist

Community

а
б
в
е

а
б
ц
к
е
Harder mapping

Linguist

\[ a \rightarrow а \]
\[ b \rightarrow б \]
\[ c \rightarrow ц \]
\[ e \rightarrow е \]

Community

\[ а \rightarrow a \]
\[ б \rightarrow b \]
\[ ц \rightarrow c \]
\[ е \rightarrow e \]

\[ \text{с} \rightarrow ц [ts] / _i \]
Harder mapping

Linguist → Community

а → а
б → б
ц → ц
к → к
е → е

Else: c → к [k]

с → ц [ts] / _ i
Even harder mapping

Linguist

Community
Even harder mapping

Linguist

Community

acht

АЦХТ

a

а

June 25, 2008  Data Management & Archiving: Data types  37
Even harder mapping

c = [ts] = “ц”

ch = [tʃ] = “ч”
Even harder mapping

Linguist

Community

\[ c = [ts] = \text{“ч”} \]

\[ h = [h] = \text{“х”} \]

\[ ch = [tʃ] = \text{“ч”} \]
Linguist | Community
---|---
\(a\) | \(a\)
\(a\) | \(ц\)
\(h\) | \(ч\)
\(t\) | \(т\)
\(t\) | \(ч\)

\[ c = [\text{ts}] = "ц" \]
\[ h = [\text{h}] = "ч" \]
\[ ch = [\text{tʃ}] = "ч" \]
\[ t = [\text{t}] = "т" \]

Else: \[ t = [\text{tʃ}]_i = "ч" \]
• There isn’t one solution here, of course

• The linguist should be aware of how their transcription relates to an existing, or developing orthography

• The linguist should make the community aware if its choices will hinder mobilization so a resolution can be negotiated

• If an existing orthography is in wide use, the linguist should plan to map to it from the outset
• Stick to what’s in Unicode if at all possible
• If you’ve got a tough mapping, learn some regular expressions (or find a clever undergraduate)
• Get advice from people with experience
• Texts—including elicitations—that a linguist collects will be of interest to the community

• They are one of the easier things to package for community use (assuming orthographic issues are solved)

• Structure is fairly consistent across languages

• A data type where use of standards can pay quick dividends—for linguistic and community use
• Communities will probably not want interlinear text

• Parallel texts (e.g., one page documented language, facing page vehicular language) are relatively easy to create

• In revitalization/maintenance contexts, communities may need more than this
• Revisiting interlinear glossed text

Cicko, [ch’ara]’a goj, ’i bu’u.
cat.ERG fish & see.CVPAN 3S.ABS B.eat.PRS

‘The cat sees a fish and eats it.’

• A gloss is probably best understood as a highly restricted view of a dictionary entry

• A kind of mnemonic ID—i.e., it’s symbolic
Revisiting interlinear glossed text

\[ Cicko, [ch’ara \ 'a \ goj,] \ 'i \ bu’u. \]

cat.ERG  fish  &  see.CVPAN  3S.ABS  B.eat.PRS

‘The cat sees a fish and eats it.’

What is a gloss?

- A gloss is probably best understood as a highly restricted view of a dictionary entry.
- A kind of mnemonic ID—i.e., it’s symbolic.
Software that actively manages two-way relationships between glosses and lexical entries is hard to come by (FieldWorks?)

But you can try to

- Be consistent in a morpheme’s gloss
- Use a unique gloss for each morpheme
- Include a gloss “ID” in a lexical database
- A gloss that can be directly associated with a dictionary record is a powerful tool
Integration

Texts

Unique Gloss
Integration

Texts --> Unique Gloss --> Record Lookup --> Dictionary Database

Data Management & Archiving: Data types

June 25, 2008
Integration

Texts

Unique Gloss

Dictionary Record

Dictionary Database

Record Lookup
Integration

Texts

Unique Gloss

Dictionary Record

Dictionary Database

Record Lookup

Rich Applications
• Your data management plan should be integrated with your data mobilization plan

• Community mobilization may require

• Extra content to be added—and an extended conceptual model (e.g., for a reverse dictionary)

• Choice of one possible implementation over another (e.g., a readily mappable transcription system)

• Or both (e.g., gloss as ID)