Toward interoperability and a cyberinfrastructure for language data

Helen Aristar-Dry
LINGUIST List & Eastern Michigan U.
Technology for language research
  Promise vs reality
  Reasons for the gap:
    moving from a static (paper) infrastructure to dynamic, distributed infrastructure requiring interoperation
  Progress: in standards needed for shallow interoperation
  Needed: standards development & use for deep interoperation
Medieval Help Desk

- From a Norwegian TV show "Øystein og jeg" on NRK
- Posted to YouTube at least 3 times. The Feb 27, 2007 posting had 56,234 views by July.
- Google found 1,390,000 references to it

Why is this video so popular?
Every language-oriented field is currently in a time of technological transition. This is certainly true of documentary linguistics.

Reflect: what technology was not available 20 years ago? What was fieldwork like?

Example: Biao Mien notecards (1982)
http://emeld.org/school/classroom/images/
pæn

miən

? pwaŋ³ rhyme rogue, if cog.
Fieldwork products today: David Harrison’s Tofa data in the E-MELD School of Best Practices:
http://emeld.org/school/case/tofa/

But fieldwork, like the rest of linguistic research, is in technology transition

See: Jim Fox’s checklist for field research
http://www.stanford.edu/~popolvuh/field-checklist.htm
Data slips AND computers?

Also doesn’t go far enough: once the linguist has the material on his computer what happens?

- Can others use it?
- Can he use it in other programs?

Tofa data is on the web and in a form that makes it reusable. But this is rare.

Most data, even digital data, resides on a single computer and isn’t sharable.
Many language documenters are frustrated with new technology because there is:
- Considerable learning required
- But less ‘payoff’ than expected

A gap between the promises about digital research and the reality
- **Promise** of data re-use, mobility, interoperability
- An example of **reality**: Internet data search
Ex: Student wants info on W. Sisaala

- Searches “W. Sisaala language” via Google
- 2170 hits, but:
  - Most are on culture, not language. Omitting non-language: count = 972' 
  - Most language references refer to the same work (his own):
    - Omitting E-MELD and LINGUIST List: count = 42
    - Finds no actual data or documentation
Google Search for W. Sisaala

Does find “The Language Museum”
- Which says it has data from 2000 languages
- And an ad:
  “To experience the best "Stag Weekends" in the world, visit RedLeisure.“
  http://www.language-museum.com/

- Conclusion: a Google search has high recall, low precision
Imagine searching Google for:

**Languages U and E**

<table>
<thead>
<tr>
<th>Name</th>
<th>Code ISO-639-3</th>
<th>Alt name</th>
<th>Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>kjg</td>
<td>Khmu</td>
<td>Austro-Asiatic</td>
</tr>
<tr>
<td>U</td>
<td>uuu</td>
<td>U</td>
<td>Austro-Asiatic</td>
</tr>
<tr>
<td>U</td>
<td>bod</td>
<td>Tibetan</td>
<td>Sino-Tibetan</td>
</tr>
<tr>
<td>E</td>
<td>eee</td>
<td>Kjang E</td>
<td>Tai-Kadai</td>
</tr>
</tbody>
</table>
OLAC (Open Language Archives Community)
- “union catalog” of metadata of the holdings of 35 digital language archives
- http://linguistlist.org/olac/

BUT:
- Finds only 3 records (Ethnologue, LINGUIST List, Rosetta) – no data

WHY?
- Are there no corpora of African languages in 35 digital language archives? Maybe, maybe not.
- Don’t know, because 31 archives don’t fully comply with OLAC standard for metadata:
  - Don’t specify subject language
  - Don’t use language codes
ODIN: Online Database of InterliNear text

- A ‘linguistic spider’ or web crawler
- Developed by Will Lewis & Scott Farrar: see proceedings of 2003 E-MELD Workshop

Service currently reports more than:

- 34,000 instances of Interlinear Glossed Text examples (shortly to be augmented to 160,000)
- from over 700 different languages
- in more than 2,200 different linguistic documents

http://www.csufresno.edu/odin/

http://www.linguistics-ontology.org/odin/
## ODIN

The Online Database of Interlinear Text

List of documents and pages with Interlinear examples for **Abkhaz (ABK)**
(Alternate names and dialects for Abkhaz are Abxazo, Abzhui, Bzyb, and Samurzakan)

<table>
<thead>
<tr>
<th>URL</th>
<th>#</th>
<th>Verified</th>
<th>Raw</th>
<th>xml</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.ling.udel.edu/colin/research/papers/Disagreement.pdf">http://www.ling.udel.edu/colin/research/papers/Disagreement.pdf</a></td>
<td>1</td>
<td>Highest</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><a href="http://lib.org.by/djur/L_Languages/LF/Formal%20theory/Kahre%20%20%20Aspects%200f%20">http://lib.org.by/djur/L_Languages/LF/Formal%20theory/Kahre%20%20%20Aspects%200f%20</a>...</td>
<td>3</td>
<td>High</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><a href="http://lings.i.man.ac.uk/Info/Staff/KEB/Teaching/LI105/LI105%20Tut4.pdf">http://lings.i.man.ac.uk/Info/Staff/KEB/Teaching/LI105/LI105%20Tut4.pdf</a></td>
<td>1</td>
<td>High</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><a href="http://www.bhargav.com/books/Math/Aspects_of_Negation.pdf">http://www.bhargav.com/books/Math/Aspects_of_Negation.pdf</a></td>
<td>2</td>
<td>Auto</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><a href="http://www.uni-leipzig.de/~autotyp/download/infl.pdf">http://www.uni-leipzig.de/~autotyp/download/infl.pdf</a></td>
<td>1</td>
<td>Auto</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><a href="http://www.unipb.it/wwwling/saperinitenessNEW.pdf">http://www.unipb.it/wwwling/saperinitenessNEW.pdf</a></td>
<td>1</td>
<td>Auto</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><a href="http://www.ling.uni-osnabrueck.de/aktuelles/finitheit/Bisang-Fin1.pdf">http://www.ling.uni-osnabrueck.de/aktuelles/finitheit/Bisang-Fin1.pdf</a></td>
<td>1</td>
<td>Auto</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

### Notes:

- The results presented here represent data found on the Web that contain suspected instances of Interlinear Glossed Text (IGT). Misidentifications are possible.
- For those results that indicate **Verified** as "Highest" or "High", all instances of IGT in the document have been manually verified to be IGT and to be in the language specified. Highest verification indicates that the results suffered no corruption in the extraction process. "Auto" indicates that these data were added through an automated process without supervision. "Low" indicates that the language was not verified, although the instances discovered are IGT.
- If **Raw** or **XML** appears as "Yes", then you can view the actual instances by clicking the link. Note: instances are only
ODIN

The Online Database of Interlinear Text

The following interlinear glossed text data was extracted from a document found on the World Wide Web via a semi-automated process. The data presented here could contain corruption (degraded or missing characters), so the source document (link below) should be consulted to ensure accuracy. If you use any of the data shown here for research purposes, be sure to cite ODIN and the source document. Please use the following citation record or variant thereof:

URL: http://www.ling.udel.edu/colin/research/papers/Disagreement.pdf
(Last accessed 2005-03-29).
ODIN: http://www.csufresno.edu/odin/igt_raw.php?id=980&langua...
ODIN can only display a fraction of the actual examples of IGT it finds because:

- Ethics forbid publications of data without appropriate credit
- Most of the data can’t be properly credited

Why?

Document authors didn’t add metadata to the web pages
Although the Internet is estimated to contain:
- over 30 billion web pages
- on 100 million Internet websites

Student can’t access any actual linguistic data (on a language with over 30,000 speakers)

Compare this reality with the promised digital search scenario...
Researcher (who is speaker of W. Sisaala) logs on from her rural village in Ghana

Finds texts, audio, and video in W. Sisaala, as well as information about it

Finds numerous online scholarly reviews of the online corpora so she knows which corpus to choose

Can analyze it using ‘Scholar’s Workbench for African Languages Online’ (SWALO), combining it with multidisciplinary data using advanced analysis and visualization tools
Or she can download it by entering her URI (Universal Researcher Identifier). Download is free of charge

Of course

Download is recorded and the material’s creator automatically cited in appropriate scholarly index.

The multimedia material plays immediately on her own machine

Of course

The text is immediately available to her client search software.

Of course

If she edits, stand-off annotation is uploaded to SWAL so that the next user can see what she’s done and chat with her about it over MSN  

(Microsoft Scholar’s Network)
“The primary access to the latest [research] findings . . . is through the Web. . . .

Crucial data collections in the social sciences are online and remotely accessible

Groups collaborate across institutions and time zones, sharing data, ideas, and access without travel.

Researchers
- Analyze data remotely using advanced tools
- Visualize the results of complex data sets in new and exciting ways”
A Vision of Digital Language Resources

- **Preservable**: formats are not vulnerable to physical decay or obsolescence of hardware & software
- **Intelligible**: content is easily understood by future scholars & community members
- **Accessible**: distributed resources are easily discovered and accessed
- **Interoperable**: documentation created by different language documenters is easily searched, compared, and re-purposed
How are we doing?

- Cyberinfrastructure for communication & collaboration
  - Email
  - Also:
    - Distribution lists, RSS
    - Skype
    - Video-conferencing
    - Instant messaging
- Good cyberinfrastructure
- General information retrieval:
  - Online library catalogs
  - Online journal indexes: JSTOR, Science Direct, First Search, etc.
  - Full-text encyclopedias & reference works
  - Wikipedia, Scholarpedia, Citizendium, Glottopedia
- Pretty good cyberinfrastructure
With regard to the linguistic data & documentation:

- Data retrieval
- Data analysis
- Data comparison
- Data re-purposing

Inadequate cyberinfrastructure
Why the Gap?

- We are moving from static paper infrastructure to dynamic, collaborative cyberinfrastructure which
  - Aggregates resources from distributed sites
  - Requires interoperation

- Moreover, cyberinfrastructure for data access
  - Requires deep (not shallow) interoperation
  - Requires domain-specific standards and participation of domain experts (linguists, speakers)
Two kinds of Interoperation

- **Shallow interoperation**
  - Based on the **surface content** of plain text
  - Generic to **all** domains
  - Based on the ubiquitous HTTP infrastructure

- **Deep interoperation**
  - Based on **underlying** concepts and structures
  - Built for a **specific** domain
  - Requires a **domain-specific** infrastructure (e.g. schemas, markup, controlled vocabularies)

  --From Gary Simons’ talk at German Linguistic Society, Bielefeld, 2006
Google is based on shallow interoperation
But so are most of the mature infrastructures we know:

- Communications infrastructure (Email, etc.) is ‘generic to all problem domains’
- Infrastructure for retrieval of paper resources, e.g. search for a paper book (c. 1987):
  - Student goes to library
  - Searches card catalog
  - Finds call number (for African Languages)
An elaborate infrastructure

- Human & institutional infrastructure (author, editor, printer, publisher, librarian)
- Numerous standards allowing interoperation:
  - Shape of book (codex not scroll): took 300 years to catch on
  - Book sizes (folio, quarto)
  - Page numbers (none in Medieval books)
  - Parts of book (25 ordered parts)
Standards: 25 parts of a book

- Front cover
- Front endpaper
- Flyleaf
- Front matter
  - Frontispiece
  - Title page
  - Copyright page
  - Table of contents
  - List of figures
  - List of tables
  - Dedication
  - Acknowledgments
  - Foreword
- Preface
- Introduction
- Back matter
  - Appendix
  - Glossary
  - Index
  - Notes
  - Bibliography
  - Colophon
- Flyleaf
- Rear endpaper
- Rear cover
- Cataloguing system (Anglo-American Catalog Rules)
  - how to handle ‘Anonymous,’ titles beginning with ‘The’ and ‘A’
  - published 1908, revised in 1978
- Call numbers (Dewey Decimal or Library of Congress)
- ISBN & ISSN
- ANSI/NISO Z39.41: correct way to place info on the book spine, including the author’s name
Paper infrastructure (book standards) took almost 3000 years to develop
Paper infrastructure provides shallow interoperation

- "Generic to all problem domains"
  - Standards developed by typesetters, printers, librarians
  - None reaches very far ‘down’ into the content of the book
  - So none required development by ‘domain experts’ (e.g., language experts)

- "Based on the surface content of plain text"
- "Based on the ubiquitous ftp ‘find-the-paper’ protocol"
Standards needed to make digital data interoperable

Example: digital lexicon (text only).
  - Machine requires standardized:
    - Character encoding
    - Metadata
      - Elements (creator, contributor, access rights, etc.)
      - Values including codes, e.g. language identification
Machine also requires standardized:

- **Annotation** for linguistic description, e.g., does “P” mean ‘Past,’ ‘Perfect,’ ‘Perfective,’ or ‘Pronoun’?
- **Structure of the linguistic data type** (e.g., a lexical entry contains a headword, pronunciation, etymology, sense)
- **Structure of the Document** (lexicon schema)

- Unlike the standards for paper infrastructure, all these standards must be developed by domain experts (language documenters, language teachers, linguists)
But we’re making progress

- Internet only became a potential research tool 20 years ago
- As we have seen, the necessary standards are already being developed
  - IMDI or OLAC metadata standard for describing linguistic resources
  - Leipzig Glossing Rules, GOLD, & DatCat’s (ISO TC37 SC4) for linguistic terminology and annotation
  - Schemas for IGT & Lexicons (numerous competitors, but some metastandards: LMF, LIFT)
  - Language codes – a recent ‘success’ that shows the need for domain experts in standards development
Progress (2): Tools

- Generic software tools now:
  - Write Unicode
  - Export in XML (though not very informative XML)
- Limited number of linguistic tools are available to facilitate standards compliance, e.g., tools to write metadata (ORE, IMDI metadata creator, FieldHelper) and submit it to a search engine.
- More will be developed once we
  - Agree on and implement linguistic annotation standards
  - Develop data models and document schemas
“By analogy with the word ‘infrastructure’ ‘cyberinfrastructure’ intends to convey both its importance in . . . supporting scholarly endeavors, and the idea that you eventually will be able to use it without thought”

--Indiana U. Cyberinfrastructure Newsletter, March 1, 2007

- But right now we have to think about it.
- Hence the ‘standards’ part of this class!
Right now . . .

- Language documenters have to . . .
  - Participate in developing the digital standards required by the domain
  - Change their work habits somewhat in order to comply with the standards

- Because for language data they are at once
  - The end users
  - The data providers
  - The domain experts
Right now . . .

- They are the only ones who can implement deep interoperation over linguistic data
- And make sure their followers ‘don’t have to think about infrastructure’
Not thinking about infrastructure