Resources for Endangered Languages

Specifications for a Roadmap

Dafydd Gibbon
U Bielefeld, Europe

LREC 2004 post-conference workshop
Building the LR&E Roadmap: Joint COCOSDA and ICCWLRE Meeting

Lisbon, 30 May 2004
Overview

1. Introduction: current EL resource situation
2. Snapshot of some documentation initiatives
3. Documentation logistics: an overview
4. Workable Efficient Language Documentation
5. Some resource types
6. PSI: the Principle of Securing Interpretability
7. Accessibility: pocket metadata - import + export
8. Main goal: resources for endangered languages (1)
9. Main goal: resources for endangered languages (2)
10. Milestones
11. Recommendations
Introduction: current EL resource situation

Key concept:

language documentation (~ language resources)

Introduced by Lehmann ("language museum") and Himmelmann ("language description vs. language documentation") around 1990.

Thus: the development in general linguistics arose at about the same time as the LRE paradigm in the SAM and EAGLES projects.

Language documentation is embedded in contexts such as

- heritage preservation (texts, audio-visual recordings)
- education (primers for spelling, vocabulary; readers)
- language maintenance (large dictionaries, translation, public media)
- language revitalisation (administrative and political re-instatement)
Snapshot of some documentation initiatives

Some organisational and funding initiatives:

- **FEL:** Foundation for Endangered Languages (Nick Ostler)
- **ELF:** Endangered Languages Fund (Doug Whalen)
- **GbS:** Gesellschaft für bedrohte Sprachen (Hans-Jürgen Sasse)
- **HRELP:** Hans Rausing Endangered Languages Project (SOAS)
- **DoBeS:** Dokumentation Bedrohter Sprachen (VW Foundation)
- **E-MELD:** Electronic Metastructures for Endangered Languages Documentation (Linguist List - Tony Dry, Helen Aristar-Dry)
- **ALP:** ROSETTA 1000 Language Project / ALL Language Project (Jim Mason)

The good news:

diverse funding is available for diverse needs.

The less good news:

there is generally little interest in or awareness of state-of-the-art LRE.

Understandable, but controversial priorities:

if in doubt, *description* first, then *application*,
and LRE standard *documentation* later (if at all)
Documentation logistics: an overview
Workable Efficient Language Documentation

The WELD Four Level Content concept:
1. Primary documentation: recordings, texts.
2. Secondary documentation: annotations, lexica, sketch grammars, ...
3. Primary description: large coverage lexica, descriptive grammars
4. Secondary description: theoretical typological and formal studies

The WELD Five Procedural Criteria concept:
1. Comprehensive
2. Efficient
3. State-of-the-art
4. Affordable
5. Fair

Operationalising the concepts with technology initiatives:
LLSTI, Local Language Speech Technology Initiative
(Roger Tucker, Ksenia Shalonova)
ELSNET (Steven Krauwer and many others ...
Some resource types

<table>
<thead>
<tr>
<th>Code</th>
<th>Noun</th>
<th>Meaning</th>
<th>Code</th>
<th>Noun</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>a$j$</td>
<td>$n$</td>
<td>blood; life</td>
<td>E$dlYdlUp&quot;$</td>
<td>$n$</td>
<td>corn</td>
</tr>
<tr>
<td>e@lo$vle@</td>
<td>$n$</td>
<td>bee</td>
<td>E&quot;IU&quot;p&quot;</td>
<td>$n$</td>
<td>bean</td>
</tr>
<tr>
<td>e#one#</td>
<td>$n$</td>
<td>age</td>
<td>gbU$gb$</td>
<td>$adj$</td>
<td>red</td>
</tr>
</tbody>
</table>
PSI: the Principle of Securing Interpretability

Lexicon:
Conversion of inconsistently structured print media, undocumentated Shoebox databases into standardised formats with appropriate metadata.

Character encoding:
Conversion of proprietary - often unknown - fonts into Unicode standard character definitions, if possibly by glyph comparison of printed matter - always scan your documents too!

Interlinear glossed text:
Conversion of visually formatted text - i.e. text which is unstructured in terms of coherent text objects such as tables - into coherent text objects into an XML format

Annotated recordings:
Conversion of various received formats (esps/waves+, Transcriber, Praat) into a generic format TASX

Linguistic descriptions:
As far as possible, conversion of received English, French etc. descriptions and metadata into GOLD (General Ontology for Linguistic Descriptions)

Gibbon, Bow, Bird, Hughes, Securing Interpretability ... LREC 2004
Access: pocket metadata - input + export

CorpusMetaData 02-3-12:HanDBase Export
RecordID: Agni2002a
LANGname(s): Agni, Anyi
SILcode: ANY
Affiliation: Kwa/Tano
Lect: Indénié
Country: Côte d'Ivoire
ISO: CI
Continent: Africa
LangNote: </LANGNote>
SESSION: FieldIndoor
SessionDate: 02-3-11
SessionTime: 8:57
SessionLocale: Adaou
Domain: Syntax
Genre: Questionnaire
Part/Sex/Age: Kouamé Ama Bié f 35
Interviewers: Adouakou
Recordist: Salffner, Gibbon
Media: Laryngograph
Equipment: 1) Audio: 2 channel, l laryngograph, r Sennheiser studio mike
2) Stills: Sony digital
3) Video: Panasonic digital (illustration of techniques)
SessionNote: f. Adouakou phrases repeat

<?xml version="1.0"?>
<CorpusMetaData>
  <Record>
    <RecordID>Agni2002a</RecordID>
    <LANGnames>Agni, Anyi</LANGnames>
    <SILcode>ANY</SILcode>
    <Affiliation>Kwa/Tano</Affiliation>
    <Lect>Indénié</Lect>
    <Country>Côte d'Ivoire</Country>
    <ISO>CI</ISO>
    <Continent>Africa</Continent>
    <SESSION>FieldIndoor</SESSION>
    <SessionDate>03/11/2002</SessionDate>
    <SessionTime>08:57 am</SessionTime>
    <SessionLocale>Adaou</SessionLocale>
    <Domain>Syntax</Domain>
    <Genre>Questionnaire</Genre>
    <PartSexAge>Kouamé Ama Bié f 35</PartSexAge>
    <Interviewers>Adouakou</Interviewers>
    <Recordist>Salffner, Gibbon</Recordist>
    <Media>Laryngograph</Media>
    <Equipment>1) Audio: 2 channel, l laryngograph, r Sennheiser studio mike
2) Stills: Sony digital
3) Video: Panasonic digital (illustration of techniques)</Equipment>
    <SessionNote>f Adouakou phrases repeat</SessionNote>
  </Record>
</CorpusMetaData>
Main goal: resources for endangered languages (1)

Description: Provision of model resources for endangered languages with different typological characteristics (audio and video recordings, texts, transcriptions, annotations, sketch grammar, extended core lexicon, appropriate acquisition tool.

Target: 2010.

Justification: A number of the model descriptive / documentary ventures are under way, but most other activities do not use state-of-the-art LRE methodologies.

Obstacles: Bottlenecks connected with the "digital divide" between commercially interesting and uninteresting languages and societies and social prejudices against minority low prestige languages.
Main goal: resources for endangered languages (2)

Prerequisites: enabling approaches and technologies, such as WELD, PSI, ...
LLSTI, BLARK

Impact: For example, TTS (text-to-speech) systems can empower pre-literate members of rural communities by providing information channels parallel to traditional social channels about health, agriculture, marketing, banking, education

Evaluation: Various evaluation techniques are needed in connection with diagnostic analysis and field functionality for local and scientific communities.
Milestones

Provision of model resources with model metadata and access portals for endangered languages with different typological characteristics in the form of:

- audio and video recordings
- transcriptions and annotations
- texts
- text markup
- core lexicon
- sketch grammar
Recommendations

Technological empowerment of endangered and other minority languages by

- more work on the PSI methodology for securing interpretability
- implementation of the WELD principles of
  - Workable
  - Efficient
  - Language
  - Documentation
- requirements specification, design and implementation of
  - the BLARK for HLT
  - according to LRE guidelines (ELSNET)
- development of basic speech technology applications (LLSTI)
- provision of access to resources via metadata portals (OLAC)