Human Language Technology Resources for Less Commonly Taught Languages: Lessons Learned Toward Creation of Basic Language Resources

Heather Simpson, Christopher Cieri, Kazuaki Maeda, Kathryn Baker, Boyan Onyshkevych

Contact Author: hsimpson@ldc.upenn.edu
University of Pennsylvania
Linguistic Data Consortium
www.ldc.upenn.edu
Program goal: create HLTs for LCTLs
- especially MT, information extraction

LDC created language packs for 13 LCTLs to support technology development efforts
- Amazigh (Berber), Bengali, Hungarian, Kurdish, Pashto, Punjabi, Tamil, Tagalog, Thai, Tigrinya, Urdu, Uzbek, Yoruba
  - NMSU: Amharic, Burmese, Chechen, Guarani (Paraguay and Argentina), Maguindanao (Phillipines), Uighur (Xinjiang, China)

Language Selection Criteria
- large population of native speakers
- relatively few language resources
- expectation of some electronic text
- intentionally vary expected difficulty of LR creation
- linguistic and geographic diversity
- include some related languages
  - Bengali, Punjabi, Urdu
- make best use of existing collaborations
  - Amazigh, Hungarian
Overview of Resources

◆ Goals
  ● support research into LCTLS
  ● test porting of HLTs
  ● test resource impoverished HLT development
  ● explore interoperability of existing resources
  ● identify LCTL LR creation issues
  ● provide framework to solicit community input on LCTL LRs

◆ Principles
  ● low cost
  ● rapid turn-around
  ● accept found data, create data to fill gaps
  ● remain cognizant of LCTL work elsewhere
    ■ ENABLER, ELSNET BLARK/ELARK, EMILLE, NMSU
    ■ 15 deliverable components in LCTL packs represent 6/9 text and 4/15 text-based modules from BLARK matrices

<table>
<thead>
<tr>
<th>Task</th>
<th>Urdu</th>
<th>Thai</th>
<th>Hungarian</th>
<th>Bengali</th>
<th>Punjabi</th>
<th>Tamil</th>
<th>Yoruba</th>
</tr>
</thead>
<tbody>
<tr>
<td>News Text</td>
<td>2,000</td>
<td>2,000</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>LCTL-&gt;English News</td>
<td>130</td>
<td>130</td>
<td>130</td>
<td>130</td>
<td>130</td>
<td>130</td>
<td>130</td>
</tr>
<tr>
<td>LCTL-&gt;English Blogs</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>LCTL-&gt;English Conversation</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>English-&gt;LCTL News</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>English-&gt;LCTL Elicitation</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>English-&gt;LCTL Blogs</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>English-&gt;LCTL Phrasebook</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Lexicon</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Encoding Converter</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Sentence Segments</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Word Segments</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>POS Tagset</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>POS Tagger</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>POS Tagged Text</td>
<td>5</td>
<td>5</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Morphological Analyzer</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Morph’ly Analyzed Text</td>
<td>5</td>
<td>5</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Named Entity Tagger</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Name Transliterator</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Narrative Grammar</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
Creation of Language Packs

◆ Resource Identification
  - individual scouting
  - “Harvest Festival”
  - native speakers

◆ Monolingual Text = base
  - Identify, harvest, remove source tags, convert to standard (UTF-8), segment, and tokenize
  - LCTLs with lack of existing raw digital text, e.g. Yoruba
    - Physical collection of newspapers in Nigeria to create 45% of Monolingual Text

◆ Parallel Text
  - LCTL -> English:
    - harvested
    - translated from monolingual text
  - English -> LCTL
    - news/blog + Special Corpora
Other Resources

◆ Lexica
  - Goal was 10,000 entries, secondarily maximal token coverage over monolingual text
  - Consulted existing electronic and paper lexica
    - Normalized target form, representation
    - Added POSes and glosses

◆ Grammatical Sketch
  - Short outlines of the features of the written language
  - Target audience: REFLEX LCTL research sites, HLT developers
Conversion/Segmentation Tools

◆ Encoding Converter
  ▶ Harvested raw text -> standard representation for LCTL
  ▶ Romanization, ISO -> UTF-8

◆ Sentence Segments

◆ Tokenizer
  • Used existing research and native speaker input
  • Especially challenging for some LCTLS (Thai, Urdu)

◆ Name Transliterator
  • Rule-based + name lists found and created
Annotated Text and Taggers

Part of speech tagger and tagged text
- Settled on 30-60,000 tokens to train POS tagger
  - POS tagged text created by in-house native speakers
  - Tagger based on MALLET toolkit

Named Entity Tagger and tagged text
- Required => 100,000 tokens to train NE tagger
  - NE tagged text created by in-house native speakers
  - Tagger based on MALLET toolkit

Morphological Analyzer and tagged text
- Rule-based, developed in XFST or written in Perl/Python
- Currently developing simple MA engine based on regular expressions
User Interfaces

◆ SimpleNET Named Entity Annotation Tool
  ● Simple NE specification: MUC (?) + TIMEX2
  ● LDC tool which allows for easy tagging of Named Entities

◆ ACK: Annotation Collection Kit Interface
  ● LDC web interface accessible from any browser
    ▪ exploits rendering engines, input of browsers
    ▪ allows annotators to work remotely
    ▪ kits can be created by non-programmers
  ● Follows LDC practice of simple, custom, portable annotation interfaces
    ▪ To accommodate short-term annotation staff
  ● Allows for multiple types of annotation
    ◆ POS Tagging
    ◆ Sentence Alignment
    ◆ Adding glosses to Lexicon entries
    ◆ QC on tool output
Sample ACK Kit

Tigrinya POS kit 32

noun
pron
adv
adj
verb
card
prep
whq
ordi
neg
art
conj
other
Collaboration

- Amazigh (Berber) with IRCAM
  - Visit by two IRCAM researchers
  - Shared text resources, specialized knowledge, normalized text and provided some much-needed annotation

- Hungarian, Uzbek, Kurdish with Media Research Centre at Budapest University of Technology and Economics (BUTE)
  - BUTE team already had access to, and in some cases was already working on, the resources needed for the LCTL language packs
  - Also had access to greater pool of educated native speakers

- Yoruba with Yiwola Awoyale at LDC
  - Had created Yoruba-English dictionary
  - Consulted on representation and encoding
  - Provided source for printer newspapers, QC
  - Annotated for NE

- Tamil with Hal Schiffman, Vasu Renganathan at UPenn
  - Had created Tamil-English verb dictionary
  - Consulted on encoding conversion
Collaboration with Native Speakers

◆ Recruitment for native speaker informants
  ● Good response, but most for remote annotation
    ■ Remote annotation support possible, but limited
    ■ Annotation Collection Kit (ACK) interface
      ◆ simpler annotation for remote training

◆ Translation agencies
  ● Found agencies for all LCTLs, but for Yoruba and Berber, turn-around and cost prevented goal amount
  ● created translation specifications
    ■ pre-segmented templates to input translations
  ● quality variable, naturally
<table>
<thead>
<tr>
<th></th>
<th>Large Languages</th>
<th>Small Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Urdu</td>
<td>Thai</td>
</tr>
<tr>
<td>Mono Text</td>
<td>14,804,000</td>
<td>39,700,000</td>
</tr>
<tr>
<td>Parallel Text (L ⇒ E)</td>
<td>1,300,000</td>
<td>694,000</td>
</tr>
<tr>
<td>Parallel Text (Found)</td>
<td>947,000</td>
<td>1,496,000</td>
</tr>
<tr>
<td>Parallel Text (E ⇒ L)</td>
<td>65,000</td>
<td>65,000</td>
</tr>
<tr>
<td>Lexicon</td>
<td>26,000</td>
<td>232,000</td>
</tr>
<tr>
<td>Encoding Converter</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Sentence Segmenter</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Word Segmenter</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>POS Tagger</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>POS Tagged Text</td>
<td>5,000</td>
<td>5,000</td>
</tr>
<tr>
<td>Morphological Analyzer</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Morph-Tagged Text</td>
<td>11,000</td>
<td></td>
</tr>
<tr>
<td>NE Annotated Text</td>
<td>233,000</td>
<td>218,000</td>
</tr>
<tr>
<td>Named Entity Tagger</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Name Transliterator</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Descriptive Grammar</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
## Phase 2 Language Packs as of 05/2008

<table>
<thead>
<tr>
<th></th>
<th>Tagalog</th>
<th>Tigrinya</th>
<th>Pashto</th>
<th>Uzbek</th>
<th>Kurdish</th>
<th>Berber</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono Text</td>
<td>774,000</td>
<td>617,000</td>
<td>5,958,000</td>
<td>790,000</td>
<td>2,463,000</td>
<td>181,000</td>
</tr>
<tr>
<td>Parallel Text (L ⇒ L)</td>
<td>203,000</td>
<td>139,000</td>
<td>180,000</td>
<td>206,000</td>
<td>163,000</td>
<td>26,000</td>
</tr>
<tr>
<td>Parallel Text (E ⇒ L)</td>
<td>65,000</td>
<td>65,000</td>
<td>65,000</td>
<td>65,000</td>
<td>65,000</td>
<td>65,000</td>
</tr>
<tr>
<td>Lexicon</td>
<td>18,000</td>
<td>0</td>
<td>10,000</td>
<td>25,400</td>
<td>6,500</td>
<td>Active</td>
</tr>
<tr>
<td>Encoding Converter</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Sentence Segmenter</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Word Segmenter</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>POS Tagger</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>POS Tagged Text</td>
<td>Yes</td>
<td>Active</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Active</td>
</tr>
<tr>
<td>Morphological Analyzer</td>
<td>Yes</td>
<td>Active</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Morph-Tagged Text</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>NE Annotated Text</td>
<td>136,000</td>
<td>123,000</td>
<td>165,000</td>
<td>93,000</td>
<td>62,000</td>
<td>60,000</td>
</tr>
<tr>
<td>Named Entity Tagger</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Name Transliterator</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Active</td>
</tr>
<tr>
<td>Descriptive Grammar</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>
Conclusions

◆ Completed 13 LCTL Language packs
◆ Many challenges, many solutions

● Long-term:
  ■ support for digital text creation in LCTLS like Yoruba, Tigrinya, Amazigh
  ■ continued efforts to help standardize digital representation

● Short-term: strategies like:
  ■ providing more support for remote annotation
  ■ implementing “Harvest Festival” model to find existing resources
  ■ getting community feedback on methods and resources
Acknowledgments

- Christopher Walker: Project Manager
- Carrie Theisen: Lead Annotator
- Mike Maxwell: Proposal Author, Project Manager, External Consultant
- Bill Poser: Research Linguist
- Mark Mandel: Research Linguist
- Programming and Annotation Staff: David Graff, David Lee

Thanks to:
- Andrew McCallum and colleagues for the MALLET toolkit
- Lori Levin and colleagues for the Elicitation Corpus
- ELRA, Tony McEnery for EMILLE