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

REFLEX-LCTL

- 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
      - http://www.elda.org/blark/matrice_res_mod.php

Goals for Phase 1 Language Packs

Task	Urdu	Thai	Hungarian	Bengali	Punjabi	Tamil	Yoruba
News Text	2,000	2,000	500	500	500	500	500
LCTL->English News	130	130	130	130	130	130	130
LCTL->English Blogs	20	20	20	20	20	20	20
LCTL->English Conversation	20	20	20	20	20	20	20
English->LCTL News	40	40	40	40	40	40	40
English->LCTL Elicitation	20	20	20	20	20	20	20
English->LCTL Blogs	10	10	10	10	10	10	10
English->LCTL Phrasebook	10	10	10	10	10	10	10
Lexicon	10	10	10	10	10	10	10
Encoding Converter	X	X	X	X	X	X	X
Sentence Segmenter	X	X	X	X	X	X	X
Word Segmenter	X	X	X	X	X	X	X
POS Tagset	X	X	X	X	X	X	X
POS Tagger	X	X	X	X	X	X	X
POS Tagged Text	5	5				5	
Morphological Analyzer	X	X	X	X	X	X	X
Morph'ly Analyzed Text	5	5					
Named Entity Tagged Text	100	100	100	100	100	100	100
Named Entity Tagger	X	X	X	X	X	X	X
Name Transliterator	X	X	X	X	X	X	X
Narrative Grammar	X	X	X	X	X	X	X

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 Segmenter
- Tokenizer
  - Used existing research and native speaker input
  - Especially challenging for some LCTLs (Thai, Urdu)
- Name Transliterator
  - Rule-based + name lists found and created

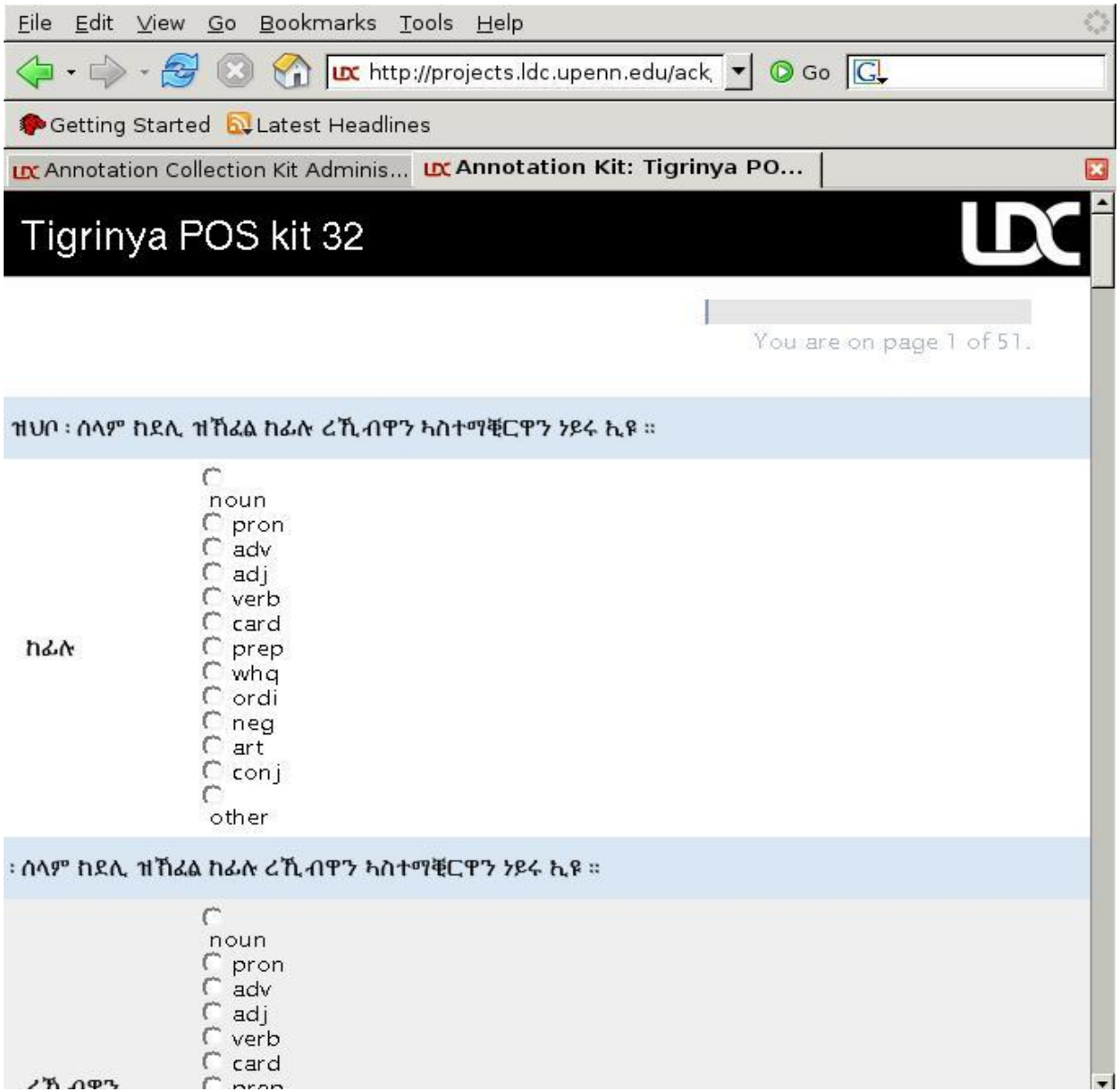
Annotated Text and Taggers

- 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



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

Phase 1 Language Packs as of 05/2008

	Large Languages		Small Languages				
	Urdu	Thai	Bengali	Tamil	Punjabi	Hungarian	Yoruba
Mono Text	14,804,000	39,700,000	2,640,000	1,112,000	13,739,000	1,414,000	363,000
Parallel Text (L ⇒ E)	1,300,000	694,000	237,000	308,000	221,000	70,000	
Parallel Text (Found)	947,000	1,496,000	243,000		230,000	2,338,000	78,600
Parallel Text (E ⇒ L)	65,000	65,000	65,000	65,000	65,000	65,000	65,000
Lexicon	26,000	232,000	482,000	10,000	108,000	182,400	128,200
Encoding Converter	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sentence Segmenter	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Word Segmenter	Yes	Yes	Yes	Yes	Yes	Yes	Yes
POS Tagger	Yes	Yes	Yes	Yes	Yes	Yes	Yes
POS Tagged Text	5,000	5,000		59,000			
Morphological Analyzer	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Morph-Tagged Text	11,000			144,000			
NE Annotated Text	233,000	218,000	138,000	132,000	157,000	269,000	189,000
Named Entity Tagger	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Name Transliterator	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Descriptive Grammar	Yes	Yes	Yes	Yes	Yes	Yes	

Phase 2 Language Packs as of 05/2008

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

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