WRITE: Comments on the Chart of LRs

Christopher Cieri
ccieri@ldc.upenn.edu
University of Pennsylvania
Linguistic Data Consortium
www.ldc.upenn.edu

REFLEX-LCTL

- Program goal: create HLTs for LCTLs
  - continues work in CallHome beginning 1995 in creating language packs
  - further inspired by surprise language experiment
  - language packs for 13 LCTLs to support technology development efforts
  - Amazigh, Bengali, Hungarian, Kurdish, Pashto, Punjabi, Tamil, Tagalog, Thai, Tigrinya, Urdu, Uzbek, Yoruba
  - Amharic, Burmese, Chechen, Guarani, Maguindanao, Uighur

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

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	

Table 1: LCTL Language Packs (Phase 1)

	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

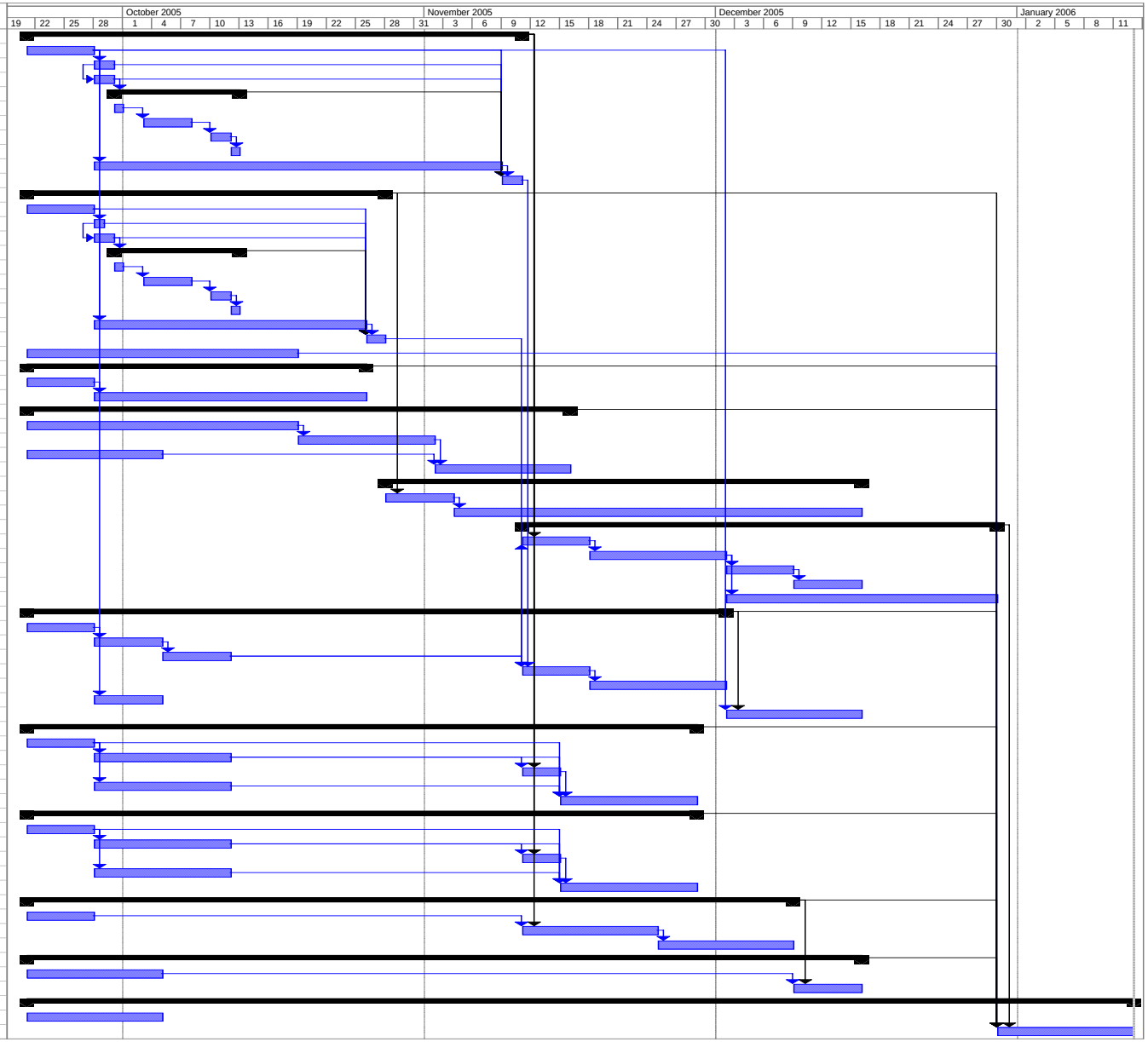
Table 2: LCTL Language Packs (Phase 2)

Dimensions & Issues

- Applications + Technologies
- Mapping between applications and technologies
  - BLARK/ELARK great start on these two and their interrelation
  - French lexicon requirements STT <> parsing
- Languages
  - official languages of the EU
  - minority & emigrant languages without official status
  - those of EU neighbors
  - allocating effort based upon language priorities
    - allocation proportionate to population = fair result?
    - low priority components, ignored, become urgent
      - need for Aceh (Acinese) on 12/26/2004
- Caution against box checking approach
  - chart better than table but …
  - When is an existing resource adequate?
    - Metadata does not make this clear? OLAC **plans** to help …
  - When to stop resource development?

Goal Workflow for Language Pack Creation

ID	Task Name	Duration
1	**Seed News Text (100k)**	**37 days**
2	Identify	5 days
3	Harvest	2 days
4	DeTag	2 days
5	**Encode**	**9 days**
6	Identify Source Encoding	1 day
7	Develop Converter	5 days
8	Test	2 days
9	Run	1 day
10	License	30 days
11	Distribute	2 days
12	**News Text (400k-1850k)**	**27 days**
13	Identify	5 days
14	Harvest	1 day
15	DeTag	2 days
16	**Encode**	**9 days**
17	Identify Source Encoding	1 day
18	Develop Converter	5 days
19	Test	2 days
20	Run	1 day
21	License	20 days
22	Distribute	2 days
23	Translate 30K News>LCTL	20 days
24	**Translate 20K Text>LCTL**	**25 days**
25	Select	5 days
26	Translate	20 days
27	**Translate 20K Elicitation Corpus>LCTL**	**40 days**
28	Develop Grammar	20 days
29	Generate Sentences	10 days
30	Enhance Tool	10 days
31	Translate	10 days
32	**Translate 175K LCTL Text>English**	**35 days**
33	Select	5 days
34	Translate	30 days
35	**Find Parallel Text**	**35 days**
36	Train BITS	5 days
37	Crawl	10 days
38	Process	5 days
39	Align	5 days
40	License	20 days
41	**Lexicon**	**52 days**
42	Identify Available Lexicons	5 days
43	Acquire	5 days
44	Format	5 days
45	Determine Coverage over Text	5 days
46	Enhance Coverage	10 days
47	Sentence Segmenter	5 days
48	Word Segmenter	10 days
49	**POS Tagged Text**	**49 days**
50	Develop Tagset	5 days
51	Develop Tagger	10 days
52	Tag Text	2 days
53	Train Human Annotators	10 days
54	Correct Tagger Output	10 days
55	**MorphAnalyzed Text**	**49 days**
56	Develop Tagset	5 days
57	Develop Tagger	10 days
58	Tag Text	2 days
59	Train Human Annotators	10 days
60	Correct Tagger Output	10 days
61	**Entity Tagged Text**	**57 days**
62	Train Human Annotators	5 days
63	Tag Text	10 days
64	Develop Tagger	10 days
65	**Name Transliterator**	**62 days**
66	Develop Transliterator	10 days
67	Test	5 days
68	**Narrative Grammar**	**82 days**
69	Write Basic Grammar	10 days
70	Check Against Collected Resources	10 days



Surprises

- The importance of making and abandoning plans
- Unicode compliance
  - bidi algorithm
  - character ordering
  - normalization forms
  - fonts
- Availability of text
  - finding Urdu translations easier than monolingual text
- Standards
  - how to write a word in Amazigh
  - choice among them
    - Gurmukhi or Shahmukhi in Punjabi
- Differences in political sensitivities
  - Tigrigna translations