

PAN Localization Project



Sarmad Hussain

*Center for Research in Urdu Language Processing
National University of Computer and Emerging Sciences
Lahore, Pakistan*

www.crupl.org

sarmad.hussain@nu.edu.pk

[Objectives]

- A regional Initiative to Develop Local Language Computing in Asia
 - To develop human resource capacity
 - To raise levels of technological support
 - To advance policy

- Focus on Developing Asian Countries

Project Information

- Funding: PAN Asia Networking, IDRC, Govt. of Canada
- Coordinator: Center for Research in Urdu Language Processing, Pakistan
- Partners:
 - ***Afghan Computer Science Association***, Afghanistan (Pashto, Dari)
 - ***BRAC University***, Bangladesh (Bengali)
 - ***Department of IT, Ministry of Information and Communication***, Bhutan (Dzongkha)
 - ***National Committee for Standardization of Khmer in Computers***, Cambodia (Khmer)
 - ***Science, Technology and Environment Agency***, Govt. of Laos (Lao)
 - ***Madan Puraskar Pustakalaya and Kathmandu University***, Nepal (Nepali)
 - ***University of Colombo School of Computing***, Sri Lanka (Sinhala, Tamil)
- Duration: 3 Years (2004 – 2007)
- Approx. 10 person team in each country (75 people in total)

Road Map to Language Computing

POLICY

1. Human Resource for NLP
 - a) CS/IT
 - b) Linguists
 - c) Computational Linguists
2. Linguistic knowledge
 1. Character set
 2. Collation
 3. Interface terminology
 4. Pronunciation
 5. Grammar
3. Language Relevant Computational Standards
 1. Character Encoding
 2. Keyboard Layout
 3. Collation
 4. Locale
4. Computational Linguistic Resources + Technology
 - a) Basic (fonts, rendering support, Input methods, locale)
 - b) Intermediate (collation, line breaking, spell checkers)
 - c) Advanced (speech and language applications, corpora and statistical NLP)

Technology Standards & Basic Applications

	Afghan-istan	Bangla-desh	Bhutan	Cambodia	Laos	Nepal	Pakistan	Sri Lanka
Standards								
Character Set	*	*	*	*	*	*	*	*
Keyboard	*		*	*	X	*	*	*
Keypad (Telephone)				X		X		
Collation Sequence	X	X	*	X	X	X	*	
Interface Terminology				X			*	
Handheld Device Interface Terminology				X		X		
Locale	*/X		*		*			
Software Applications								
Keyboard Driver	*/X		XX	*	*	*	*	*
Encoding Conversion Utility				X	X		*	*/X
Normalization and Sorting Utility	X	XX	XX	X	X	XX	*	
Find/Replace Utility		XX						
Fonts	*	*	*	*	*/X	*	*	*
Mobile/PDA Fonts	*			X		X		

X Microsoft

XX Linux

* Outside PANL10n

Technology: Advanced Applications

Lexicon	x	xx	xx	x	x	xx	*	x
Thesaurus						xx		
Natural Languages Processor		xx		x			*	
Spell Checker		xx		x	x	xx	*	
Grammar Checker				x	x	xx	*	
Text Corpus Development for NLP Applications							*	x
Text to Speech System							*	x
Speech Recognition System							*	
Machine Translation System								
Optical Character Recognition System		xx					*	x
Linux Distribution		*	xx			*/xx	*	
GNU Cash (Open Source accounting System)						xx		
	Afghanistan	Bangladesh	Bhutan	Cambodia	Laos	Nepal	Pakistan	Sri Lanka

x Microsoft

xx Linux

***** Outside PANL10n

[Activities]

- Trainings
- Standardization
- Technology Development
- Policy (where possible)

[Trainings]

■ Regional

- “Fundamentals of Local Language Computing” (Jan. 2004)
 - 60 participants from 15 countries
- “From Localization to Language Processing” (June 2005)
 - 45 participants from 12 countries
- *Afghanistan, Bangladesh, Bhutan, Cambodia, China, India, Iran, Japan, Laos, Nepal, Mongolia, Myanmar, Pakistan, Sri Lanka, Thailand*

■ National

- Afghanistan, Bhutan, Cambodia, Laos, Nepal, Sri Lanka

[Other Outputs]

- Status of Local Language Computing in Asia 2005
- Local Language Computing Network
- Local Language Computing Resources
- Handbook of Asian Language Computing

