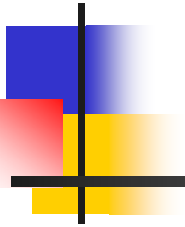


PAN Localization Phase II: Bangladesh Component



Mumit Khan
Center for Research on Bangla
Language Processing (CRBLP)
BRAC University



Highlights

- Work on all aspects - Content, Technology and Training
- Local partnerships:
 - Center for Research on Bangla Language Processing, BRAC University
 - Development Research Network (D.Net)
 - Department of Linguistics, The University of Dhaka



Phase I highlights

- Research center dedicated to Bangla language processing
- Technology development objectives met; in general exceeded original goals
- Human resource development on target
 - Gender balance not satisfactory
- Groundwork established for local and regional collaboration
- Challenges in affecting national policy



Content

- Research question: how well does OCR perform compared with manual entry?
- D.Net is the content partner
- Language resource as a side effect
- Licensing?
- More on this from Ananya Raihan (D.Net)



Training

- D.Net is the training partner
- More on this from Ananya Raihan (D.Net)



Language Technology

- BRAC University and Dhaka University
- Deliverables:
 - Phase I continuation; OCR testing and enhancements; transfer of OCR to Nepal
 - TTS application; SMS to speech application
 - Lexical resources - Corpus and WordNet



Language Resources

- Speech corpus for TTS
- Text corpus - 5M words
 - 10k manually tagged, 5M auto tagged
- Bangla WordNet - 5k entries
- Regional deliverables
 - Parallel text corpus - 100k
 - IDN's (gTLD and ccTLD)



Research Questions

3. What language and linguistic resources are available for the project languages and how difficult is it to develop these resources for a language, given the current linguistics and technical capacity?
4. What mechanisms are required for maturing available local language technology to be effectively deployed to the end-users? Which technology has larger acceptability and applicability for end-users, and thus how should language technology development be prioritized?



HR capacity development

- Department of Linguistics, DU will provide the linguistic expertise
- Part-time MA in Linguistics being developed for project staff
- Regional training opportunities (ADD workshops, etc)



Phase I boost

- Project maturity and lessons learned
- Corpus
 - Experience from news corpus
- WordNet
 - Experience from tagged lexicon
- Speech Technology
 - Experience from simple TTS



Program sustainability

- Exploring other sources of funding
 - Corporate (Microsoft, Google, ...), Public (Govt of Bangladesh), NGO (BRAC, ...), Social entrepreneur (Benetech, ...), ...
- Local collaborations - D.Net, Dhaka University, open source networks, ...
- Distributed leadership



Expected challenges

- Technological
 - *Would OCR work as well as we claim?*
 - *Linguistic knowledge needed for WordNet?*
- Policy
 - *Would the localization policy be toothless?*
- IP
 - *Can we “open content” the corpus?*