

# Status and Challenges of Local Language Computing in Bangladesh

Mumit Khan  
BRAC University

# Introduction

- Bangla/Bengali is spoken by 210 million people -- the 4<sup>th</sup> by population (*Ethnologue*)
  - Eastern-most branch of Indo-Iranian family
- National language of Bangladesh and the state language of West Bengal of India
- Rich literary history
- Very little computational linguistics work
- No substantial local language applications
- Policy, standardization, human resource issues

# Current Activities

- Localization/Local Language Computing
  - Applications: Spelling checker, OCR, Text to Speech, Speech Recognition, etc.
  - Content, Semantic search, ...
- Development Informatics
  - E-Governance, E-commerce, E-learning
  - Education: Using Biometrics to track student attendance, distance learning, ...
  - Public Health: Surveillance, reporting, ...

# Localization (L10n)

- Objective: Development of tools to that allow computing in Bangla.
- Applications: Spelling checker, OCR, Text to Speech, Speech Recognition, etc.
- Basic research: Computational Linguistics, resource development (on-line lexicon, dictionary and thesaurus, corpus, ...)
- Open Source
- Partially funded through the PAN Localization Project of IDRC, Ottawa, Canada.

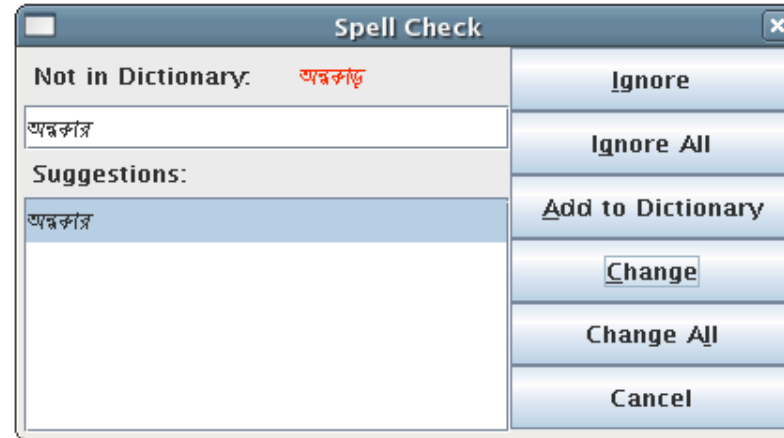
# Activity Status

- Lexicon: 100 k word list, tagging starting
- Morphology: verb morphology done, noun ongoing
- Rich-text editor with spelling checker: Done
- OCR: alpha quality; many challenges: segmentation, recognition of clusters and multi-fonts, performance, ...
- TTS: Just beginning
- ASR, MT, ... -- yet to begin

## আবেদন

শাহরিয়ার হোসেন পাভেল

একটা অন্ধকার, তার রং কালো,  
সে গ্রাস করে নিতে চায়  
বিঘত শব অর্জন, মুছে দিতে চায়  
গন্তব্য পথরেখা, সেই অন্ধকাড়, যা দিয়েছে  
মানবের মাঝে ভয় ভিত্তির জোগান,  
সেই অন্ধকার জার থেকে বাচতেই যেন  
মানুস পেয়েছে আলোড় দিষা  
ঐ অন্ধকারের কবলে পড়ে  
মানুশ বর্বর হয়েছে, উদভ্রান্ত হয়েছে,  
বিদিশায় কান্ডগ্ৰনহীন হয়েছে  
সেরকমই এক অন্ধকার যেন  
ধেয়ে আসছে, যা গ্রাষ করে নিবে  
আমাদের সোনার বাংলার স্বপ্ন,  
আমরা হবো ব্যর্থ, তাই সময় আগত  
সোজ্জার হওয়ার, রুখে দারাও অন্ধকার,  
জাগরন ঘটাও সোনার বাংলার



# Development Informatics

- E-governance and E-learning is the primary market
- Build awareness: create ICT4D courseware for Masters in Development Studies and Masters in Governance
- Public health surveillance
- Example: Developing biometric-aided IT solutions for the primary and secondary school sector (similar project in Karnataka, India)

# Policy and Standardization

- National ICT policy -- it's on paper, but can it be implemented?
  - Financial and human resources
  - Political will
- Standardization of localization-enablers can be a roadblock.
  - Terminology, keyboard, collation, etc standards
  - Distribute and let existing practice become de-facto standard?

# Human Resource Development

- HR capacity development is a key focus
  - Attract fresh graduates to the field
  - Train and retain to build capacity
- Commercial market of localized software a key enabler
  - Mobile technologies
  - Government policies (e-government for example)

# Major Challenges

- Linguistic resources
- Computational linguistics
- Human resources -- attraction, training, retention
- Market for localized software
- Significant investment needed for the long run to real applications
- Build awareness among policy makers