
NLP in Developing Asia: Lessons from PAN Localization Project



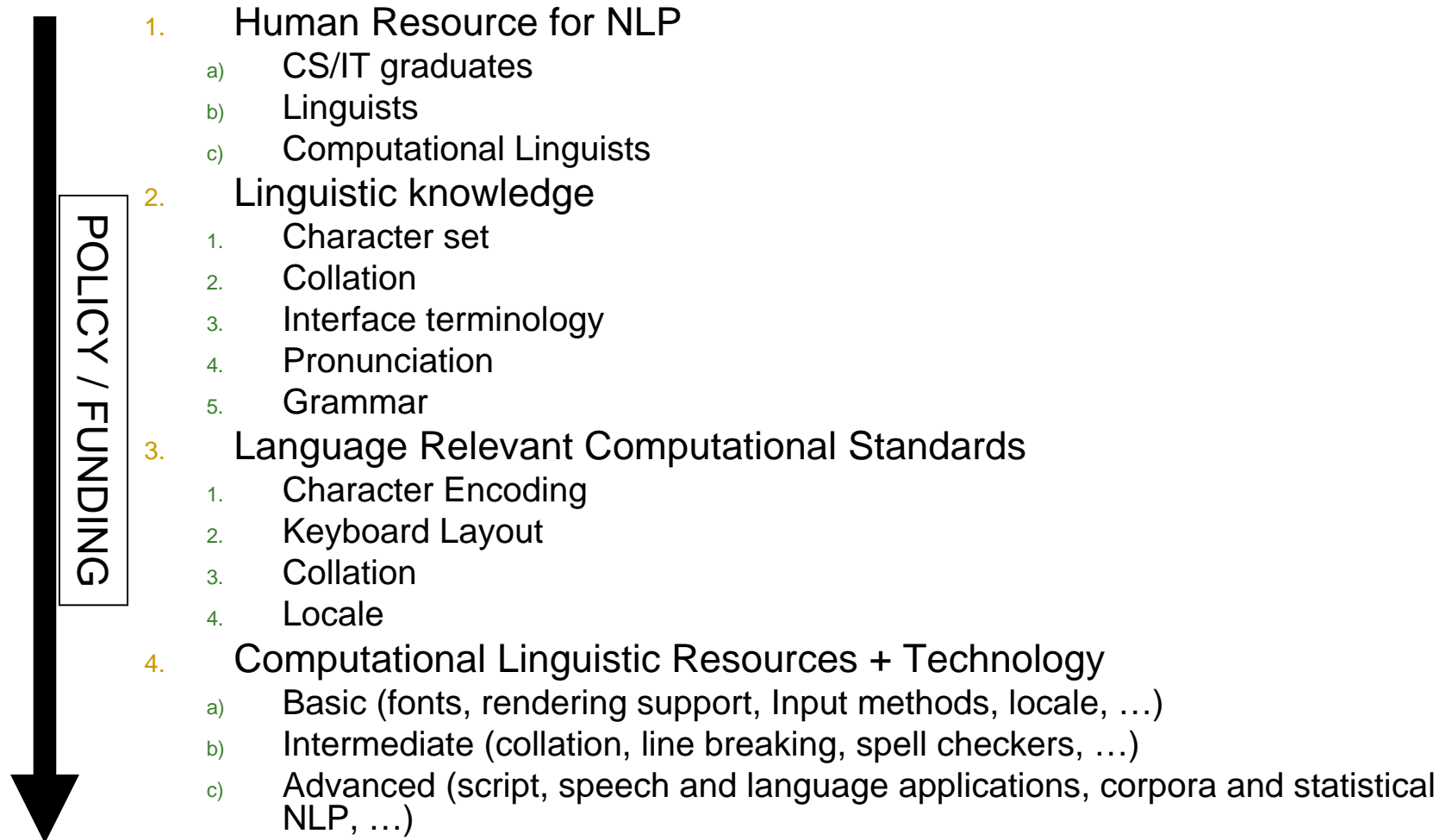
Sarmad Hussain

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

www.crup.org

sarmad.hussain@nu.edu.pk

Road Map to NLP



Status of Developing Asia: Human Resource

- Afghanistan has effectively no locally produced specialized resources
- Bhutan graduates 20 people a year with CS/IT degree. There is a quota for each ministry every year. PAN Localization had budgets to hire 5 people, but only 1 person was available
- Pakistan does not have a single linguistics degree program in the whole country
- Most graduates of CS/IT programs in Laos do not know how to program in C/C++

Status of Developing Asia: Human Resource

- Even where linguistics programs exist, descriptive linguistics is normally taught. Related courses. Also, some related courses like acoustic phonetics, are not taught at all
- There are no computational linguistics programs in almost all of developing Asia, including Afghanistan, Bangladesh, Bhutan, Cambodia, Laos, Mongolia, Myanmar, Nepal, Sri Lanka...
- Where good resources exist, there is great demand, making them difficult to retain
- Very little interest in NLP as it does not offer a career at these countries

Status of Developing Asia: Linguistic Knowledge

- In many Asian languages, basic character set for a language is still debated. Character-set for Urdu was standardized in Pakistan in 2000. Still not done for Pashto across borders.
- Lexicographic ordering is normally based on old dictionaries published through Language Authorities, and limited guidelines available for new words. Khmer collation is based on Choun Nat dictionary. Very complicated set of rules exist for Dzongkha word ordering which are still being finalized
- Terminology for computer interface does not exist in most of these languages. Terminology for Urdu published last month by Govt. of Pakistan
- Almost no work in pronunciation is done, which can be used for speech applications
- Very limited phonological, morphological and syntactic work done for most of these languages, which is mostly descriptive
- Almost no semantic work has been done for any of these languages

Status of Developing Asia: Linguistic and Computational Standards

- Minimal national standards
- Multiple ad hoc, vendor-driven conventions
- National standardization process or bodies are not well defined
- International standards still do not support many Asian languages.
 - Many languages added as recently as Unicode 3.x and later versions
 - Locale standard, CLDR 1.3 still does not list many of these Asian languages
 - Hardly any international collation standards published
 - Multiple Keyboard layouts exist, without any consensus
- In many cases, International Standards are not accepted, and ad hoc or national standards are used, e.g. Unicode is still not used widely in many of these countries

Status of Developing Asia: Basic Applications

- Platforms still do not support many of these languages.
 - Microsoft recently added support for Sinhala, Burmese, Dzongkha, Khmer, Urdu, ...
 - Many of these languages are still not properly rendered on Linux platform, Gnome, KDE, Open Office
- Locales and local language interfaces are also not incorporated

Status of Developing Asia: Intermediate Applications

- Other applications are almost non-existent :
 - Collation
 - Line breaking
 - Spell checking
- This implies limited basic word processing and web publishing for most of these languages

Status of Developing Asia: Advanced Applications

Status of Developing Asia: Policy

- Many relevant decision makers not aware of NLP
- As recent as early 2000's many governments did not even realize this as a need
 - E-Govt. programs planned without local language computing components
- Where policy was put in, no action plans were formalized

What can be done?

- National and International Policy for:
 - HR Development
 - Linguistics
 - Standardization
 - Technology and Resource Development

HR Development: Training

- Short-term national trainings
 - Detailed on a single advanced topic by a single trainer
 - Frequent, initiated mostly by national institutions and organizations
 - PAN L10n trainings in Afghanistan, Nepal, Laos, Sri Lanka
- Long-term national trainings
 - Detailed on multiple (intermediate level) topics by a single trainer, for countries really behind
 - PAN L10n placements in Bhutan, Cambodia, Laos
 - Very effective for untrained HR (experts not available for longer duration)
- Short-term regional trainings
 - Overview multiple topics by multiple trainers
 - Frequent
 - SEISA, AOSS, MLIT, PAN L10n ...
 - Limited effectiveness for untrained HR

HR Development: Training

- Long-term regional trainings
 - Detailed on multiple (intermediate to advanced) topics by multiple trainers
 - VERY BENEFICIAL, REQUIRED BUT NOT OFFERED
 - E.G. SUMMER SCHOOL IN ASIAN LANGUAGE PROCESSING (by AFNLP?)

- Develop Regional Support Network
 - to avoid duplication of effort
 - Provide development and other support
 - e.g. PAN Localization, AOSS, FOSSAP

HR Development: Degree Programs

- National focus on developing Linguistic programs
 - No faculty or funding
- National focus on developing Computational Linguistic programs
 - No faculty or funding
 - Scholarships for Developing Asian students to study NLP in other countries
 - ASIAN MASTERS IN NLP (AFNLP?)

Technology

- Develop regional resources/frameworks
 - Open Source Software
 - Open Language Resources
- Develop local language technology
 - National level priority plans
 - Support (long-term) teams to develop the technology

Policy

- Educate policy makers
 - Special trainings for policy makers
 - Role of regional organizations, like AFNLP?
- Produce literature on local language computing policy guidelines
 - E.g. to document how good policies have worked
- Help develop policy for
 - National/State Languages
 - Other Less Computerized Languages

Who will organize this work?

- PAN IDRC (Canada)
 - CICC (Japan)
 - NECTEC (Thailand)
 - APDIP UNDP
 - Korea, China, others ?
-
- Role of regional organizations like AFNLP?

Thank You