



# Asian Language Resource Network

---

January 2007  
Pan-Localization Forum  
Thimphu, Bhutan  
HASIDA Koiti (GSK) and  
KODAMA Shigeaki (NUT)



# Structure

- Sponsors
  - MEXT (Ministry of Education, Culture, Sports, Science and Technology)
  - JST (Japan Science and Technology Agency)
- Participants
  - NUT (Nagaoka University of Technology)
  - GSK (Gengo Shigen Kyokai; language resource association)
  - Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies

# Objectives

- Creating a network among Asian organizations concerning IT, in particular natural-language processing.
- Assisting them to produce/introduce new language technologies.
- Creating a text archive from the web as a basis for producing new technologies for text processing of Asian languages.
- Creating dictionaries of Asian languages from the collected texts.

# Activities

- Archiving online texts of Asian domains
- Developing tools for analyzing the archived texts
- Assessing analysis results
- Developing multilingual dictionaries
- Holding conferences on the project
- Establishing protocols and standards for language resources
- Supporting IT developers in Asia

# Survey of Asian Languages on the Web



- distribution of web pages per domain
  - distribution of languages on the web
  - multilingualism on the web
  - scripts and encodings on the Web
- etc...

# Collection of Web Pages

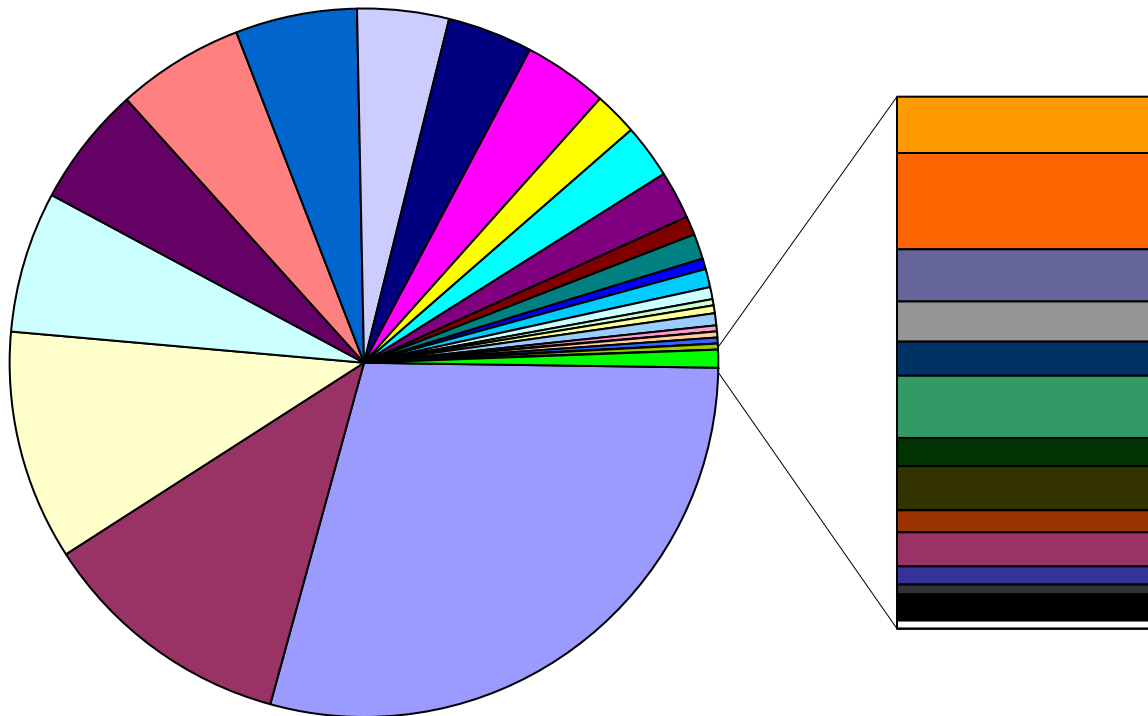
- Target regions
  - Asian ccTLDs (country-code Top-Level Domains) except for Korea, China, Taiwan and Japan
- 107,141,679 pages collected:
  - 90% of Yahoo!
  - 35% of Google

# Automatic Language Identification

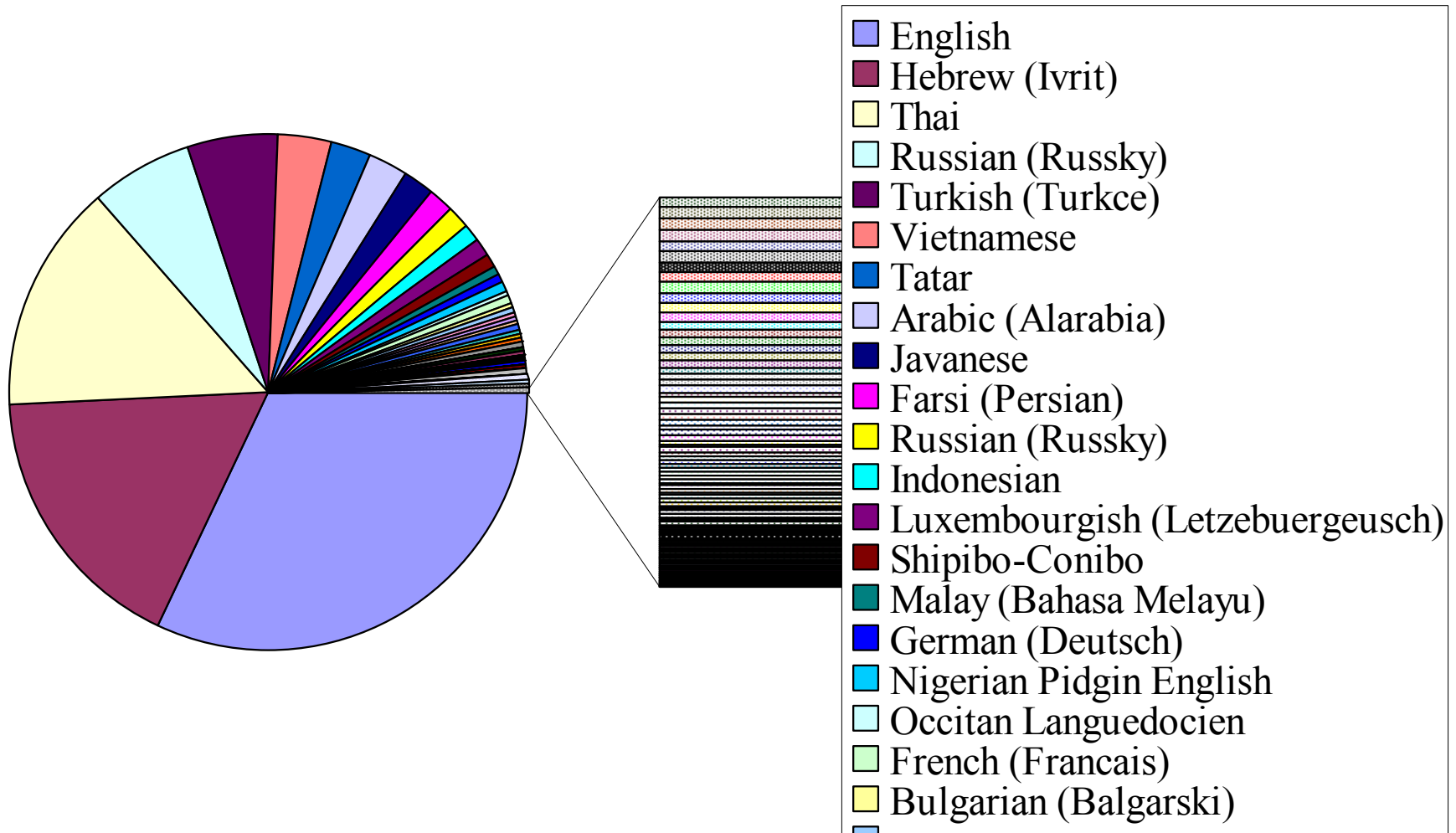


- Language/script/encoding can be automatically identified in conjunction.
- Identification targets are 101 pre-registered language/script/encoding tuples.

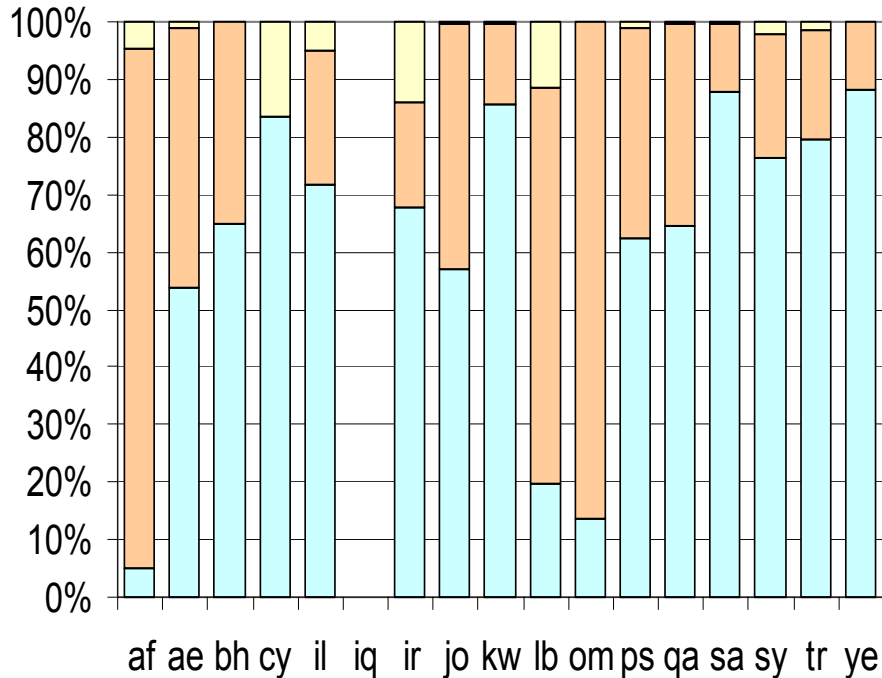
# Distribution of Domains over Asian Web Pages



# Distribution of Languages over Asian Web Pages

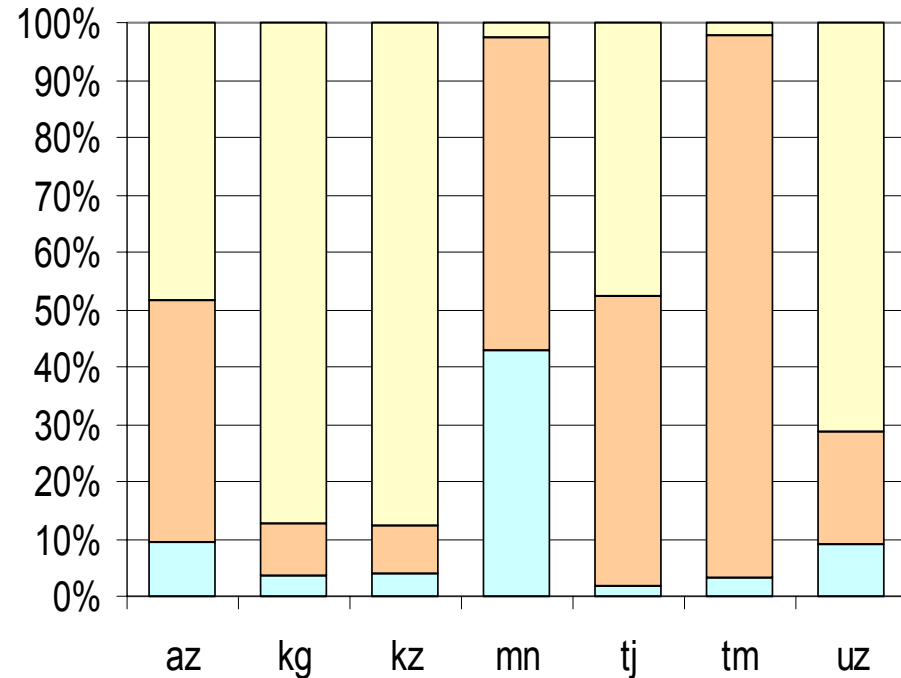


# Cross-Border Languages and Their Dominance (1/2)



**West Asia**

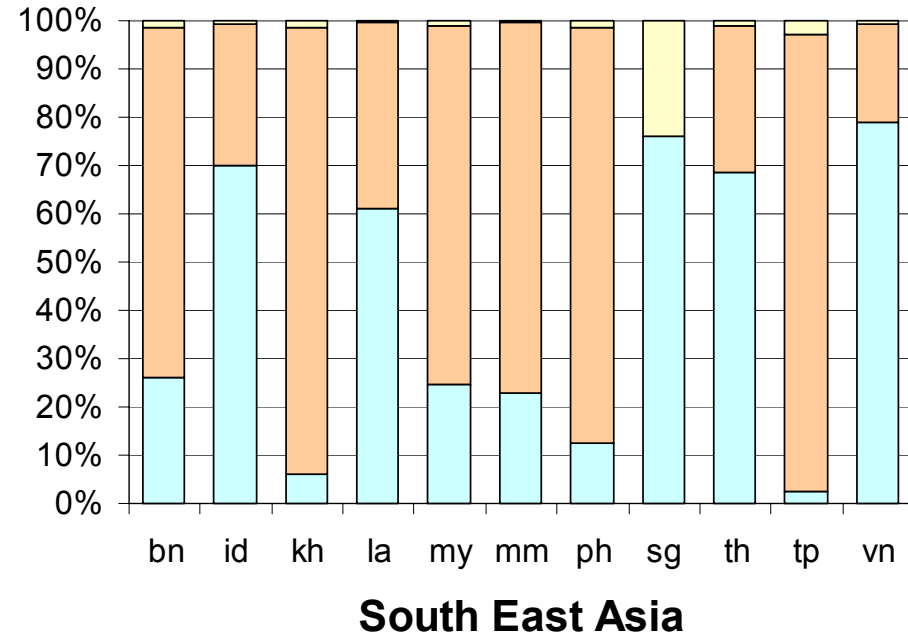
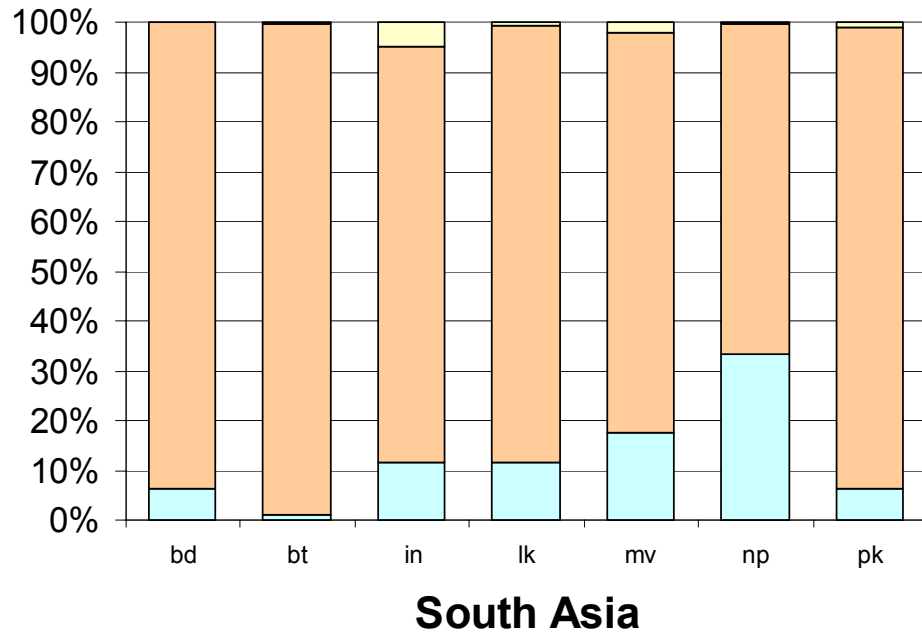
□ % Local Languages □ % English □ % Other Cross Boader Languages



**Central Asia**

□ % Local Languages □ % English □ % Other Cross Boader Languages

# Cross-Border Languages and Their Dominance (2/2)



■ % Local Languages ■ % English ■ % Other Cross Border Languages

■ % Local Languages ■ % English ■ % Other Cross Border Languages

# Multiple Encodings

Language	UTF-8 encoded documents	Otherwise encoded documents	Examples of other encodings
Vietnamese	1,934,392 (96.4%)	72,077 ( 3.6%)	TCVN, VIQR, VPS
Mongolian	48,834 (95.5%)	2,300 ( 4.5%)	Latin-Cyrillic
Hindi, Bhojpuri, Magahi, Marathi, Nepali, Sanskrit, Tamang	81,800 (78.4%)	22,544 (21.6%)	Agra, Arjun, Kiran, Kruti, Hungama, Naidunia, Shivaji, Shree, Shusha
Sinhala	4,793 (44.5%)	5,977 (55.5%)	Metta, Kaputa
Arabic	400,933 (24.0%)	1,270,189 (76.0%)	Latin-Arabic
Telugu	178 (16.6%)	894 (83.4%)	Shree, TLH
Tamil	566 (14.9%)	3,232 (85.1%)	Amudham, Kumudam, Shree, Vikatan
Hebrew	1,468,344 (12.3%)	10,488,970 (87.7%)	Latin-Hebrew
Thai	207,901 ( 2.7%)	7,544,884 (97.3%)	TIS 620
Burmese	24 ( 0.7%)	3,261 (99.3%)	WinResearcher
Turkish	20,591 ( 0.5%)	3,938,737 (99.5%)	Latin-Turkish

# Future Works

- More sample data for language identification
- More comprehensive collection
- Extending the target to gTLDs (generic Top-Level Domain, such as .com and .org)
- More exhaustive assessment of the collection/analysis results
- Extending networks over more people who are interested
- etc.

# Asian Language Resources Workshop



- Language Resources
  - corpora: text, speech, and multimedia.
  - dictionaries: monolingual, multilingual, and multimodal.
  - software tools: analysis, annotation, authoring, etc.
- 2007-03-01/02 in Akihabara, Tokyo.
- Participants overlapping with this consultation.
- Discuss how to collaborate on LRs.
- Sign on a MOU.

# Issues of Asian LRs

- requirements for LRs
  - What kind of LRs do we need for what kind of purposes?
  - How can we use them for those purposes?
  - What kind of standardization concerning LRs is needed?
  - roadmap of Asian LRs
- development of LRs
  - What kind of LRs have been or are being developed?
  - What do we need in order to improve our LRs?
  - How can we exchange knowledge between users and developers of LRs?
  - How can we promote technology transfer to countries planning to develop LRs?
- sharing LRs
  - How can we share our LRs with each other?
  - How can we dissolve relevant legal issues such as about copyrights?
  - How can we share knowledge on Asian LRs? Compile and maintain their catalogue?