Joint UNESCO and ITU
Global Symposium on Promoting the Multilingual Internet

Measuring Linguistic Diversity Through The Language Observatory

Prof. Dr. Ahmad Zaki Abu Bakar

Language Observatory Project

Universiti Teknologi Malaysia

Geneva, 9-11 May 2006

Questions to Ponder

- How do we measure linguistic diversity on the Internet?

- How does the **Language Observatory** based at Nagaoka University of Technology, Japan measure linguistic diversity?

Lead Questions on Measuring Language Diversity

ITU-T

- What are the most important indicators for measuring multilingualism in cyberspace?
- What measures are needed to build an internationally comparable indicator system?
- Who should be the main actors in data collection and analysis?
- What measures are needed to build statistical capacity-building at country level?
  - From Measuring & monitoring language diversity at Multilingualism for Cultural Diversity and Participation for All in Cyberspace at Barmako, Mali on 5-6 May 2005.

LANGUAGE OBSERVATORY

Multilingual IT Implementation

ITU-T

- Input & Output System
  - Voice
    - Voice recognition
    - Voice Synthesis
  - Text
    - **Character recognition**
    - Keyboard Layouts
      - **QWERTY / Localized keyboards**
  - Multimedia
- Knowledge Representation
- Text Processing
  - Search Engines
- Etc… the list goes on

IT Standards For MLIT. E-Business, Etc

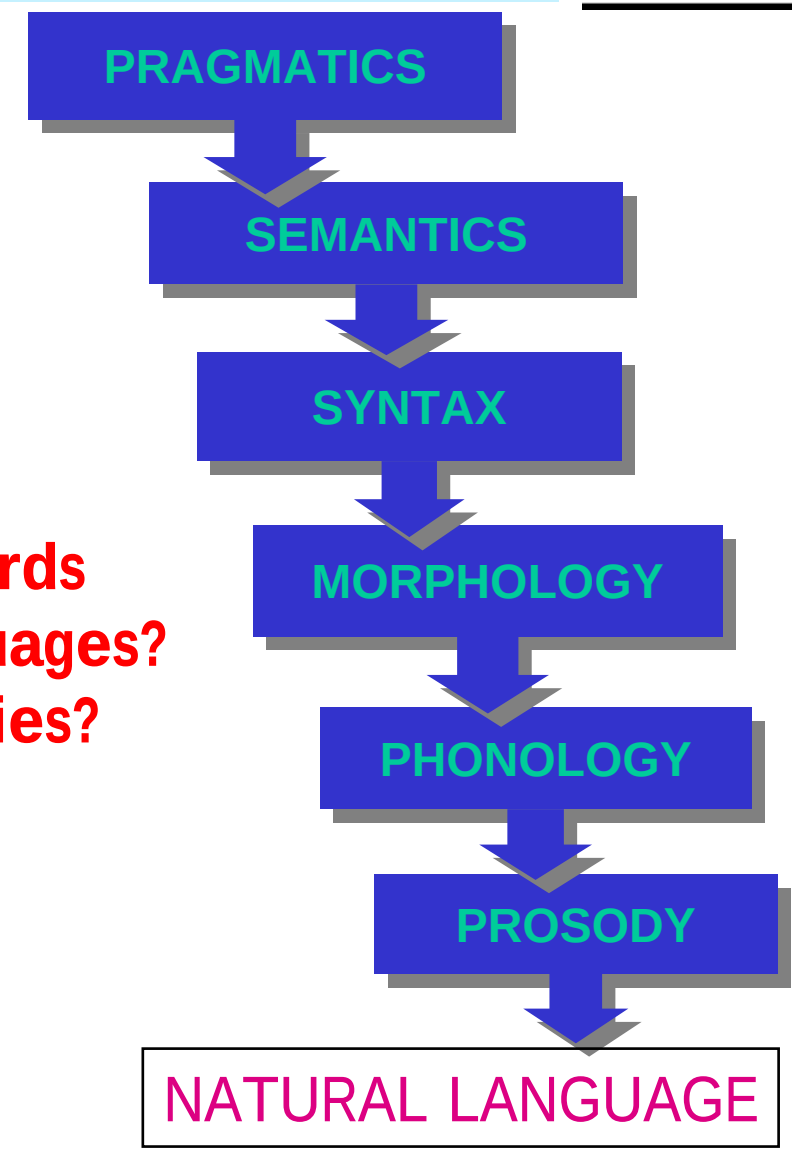
LANGUAGE OBSERVATORY

Multilingual Linguistic Issues

ITU-T

- Language Writing Systems
- Language Scripts
- Character Coding
  - ASCII / Extended ASCII
  - UNICODE / ISO 1046
- Character Sets
  - Fonts & Glyphs
- Dialects
- Computational Linguistics
- Terminology
- Etc… the list goes on

Standards For Languages? & Policies?



LANGUAGE OBSERVATORY

Diversity of World Languages

ITU-T

- UNESCO celebrates International Mother Language Day (21 Feb/Year): "to honor **6,000** Mother Languages currently spoken in the world"
- ISO 639 gives **440** language identifiers
- Universal Declaration of Human Rights (UDHR) is translated into **327** world languages
- Windows-XP handles **71** languages*
- **Google** returns pages written in **35** languages*
- Unicode 4.1 only defines 96,000 characters

LANGUAGE OBSERVATORY

note: *language count is based on UDHR language names.

Language Endangerment

- Many languages are in danger of extinction in the physical world and virtual world
- Many languages are not even born on the Internet although widely used orally
- Insufficient Multilingual Info-structure
  - More than 80 known languages character sets are still not incorporated in Unicode
    - Script Encoding Initiative
- Literacy and IT Literacy
- Digital Language Divide

Language Observatory Project - Initial Research Questions

- Language Observatory tries to catch subtle messages of less spoken languages, as far as they appear in the virtual universe, and provides such information as:
  - **How many languages are found in the virtual universe?**
  - **Which languages are missing in the virtual universe?**
  - **How many web pages are written in any given language?**
  - **How many web pages are written using e.g. Arabic script?**
  - **What kind of character encoding schemes (CES) are employed to encode a given language, e.g. Khmer?**
  - **Is there any de facto CES standard for any given language?**
  - **How quickly is ISO/IEC10646 or Unicode spreading and replacing conventional single byte encoding schemes on the net?**
- **NOTE**: the **Language Observatory is not a search engine, but an observation instrument**.

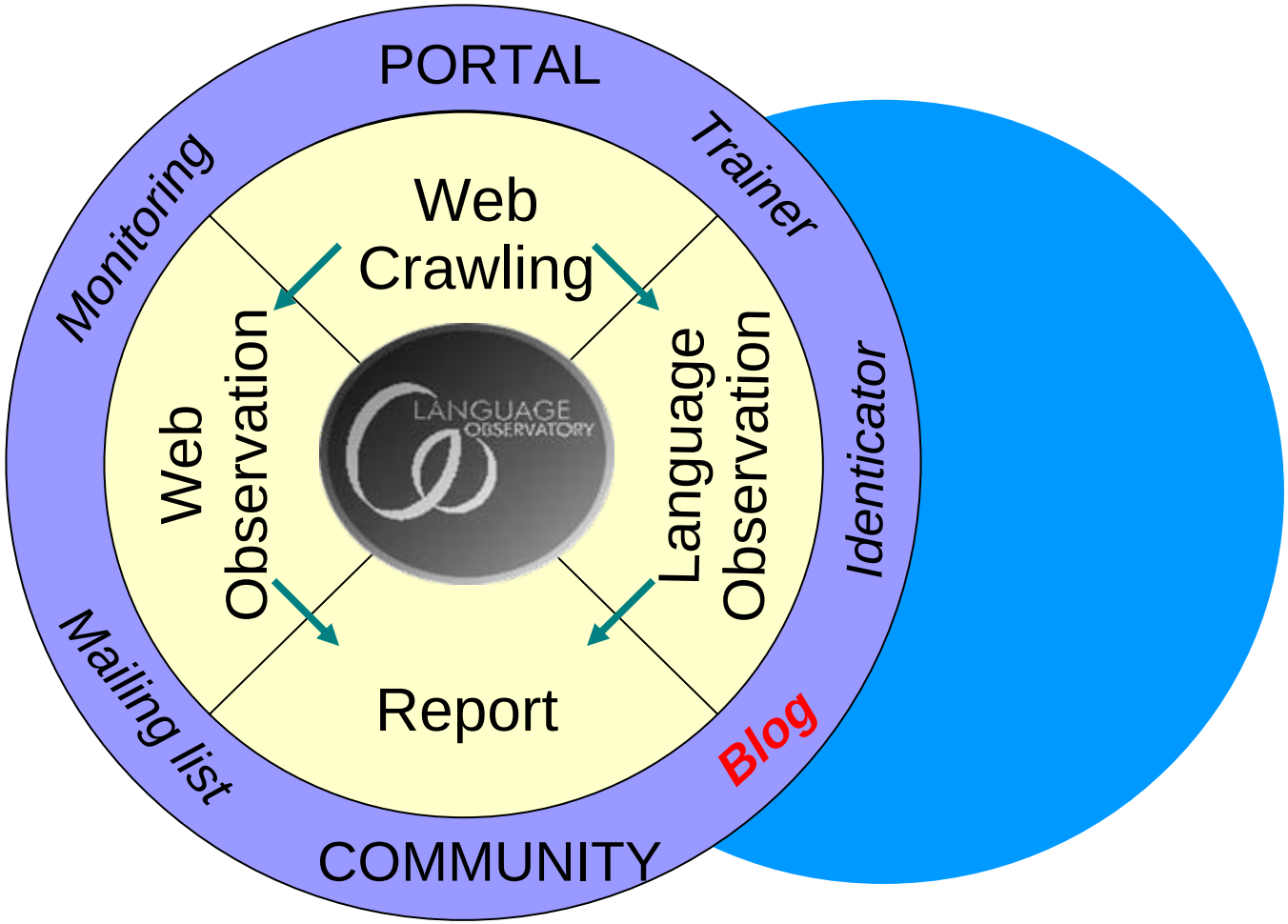
Language Observatory Project

Members

- Nagaoka University of Technology (NUT)
- Tokyo University of Foreign Studies (TUFS)
- Keio University
- *Universiti Teknologi Malaysia* (UTM)
- *Universita Degli Studi di Milano* (USM)
- Miskolc University, Hungary
- Thai Computational Linguistics Laboratory (TCL), National Institute of ICT (NICT), Thailand
- Technology Development of Indian Languages (TDIL), Department of IT, India
- Supported by Japan Science & Technology Agency & UNESCO

Language Observatory
Operations

ITU-T



For More Info Please Look At:
www.language-observatory.org

What Kind of Reports Are Produced?

ITU-T

- The Language Observatory regularly publishes the following 3 reports.
- **1. The Cyber Census Report:**
  - This report reveals the activity level of all the languages observable on the web pages of the Internet.
  - It contains statistics on;
    - number of pages and bytes by language
    - number of pages and bytes by script
    - number of pages and bytes by character encoding scheme (CES)
    - relative share in the cyberspace by language

What Kind of Reports Are Produced? Cont..

ITU-T

2. The CES Report:

- The report describes what kind of character encoding scheme (CES) is employed to encode each languages and scripts.
- Although most pages on the Internet are encoded by widely known CES, some pages are encoded by locally developed "fonts" to represent local languages.
- These locally developed fonts are not just a font, but a kind of implicit character encoding scheme.

3. The Corpus Statistics Report:

- The report gives various statistics of text data. Statistics includes;
  - Single byte/character distribution for given CES text
  - Two bytes/characters distribution for given CES text
  - Three bytes/characters distribution for given CES text
  - Single word distribution for given language/script text

Relevance of the Reports

ITU-T

- To reveal actual usage of languages in the virtual universe
  - to assist policy makers, international organizations and governments to understand the real situation of unbalanced usage of languages in cyberspace.
- To assist ICT developers and policy makers to understand the technical problems behind "Digital Language Divide".
  - To enable Technical specifications of observed Implicit CES be developed to assist development of converters.

LANGUAGE OBSERVATORY

Where Is Unicode Used?
Asian & African Summary

ITU-T

	Asia except CJK	Africa
# of ccTLDs	46	60
date of survey	September 2005	December 2005
# of domains (servers) visited	33,694	182,671
# of domains with at least one Unicode page	5,008 (15%)	9,184 (5%)
# of unique pages collected	43 million	77 million
# of pages using Unicode	5.6 million (13%)	6.6 million (11%)

LANGUAGE OBSERVATORY

Where Is Unicode Used?
by ccTLD, Asian Domains

ITU-T

Unicode documents on web servers in Asian TLDs (1)					
TLD	Unicode docs (%)	Unicode domains (%)	TLD	Unicode docs (%)	Unicode domains (%)
ae	38.4	18.6	ir	55.4	64.3
af	49.6	10.3	jo	30.1	10.1
az	18.8	34.2	kg	9.5	0.8
bd	51.1	8.3	kh	1.7	5.6
bh	0.6	6.9	kw	4.3	17.2
bt	5.4	15.4	kz	14.5	11.4
cy	5.0	17.9	la	0.2	5.4
id	6.3	13.3	lb	14.7	17.7
il	5.9	11.4	lk	31.3	19.5
in	31.2	24.6	mm	0.2	8.7

LANGUAGE OBSERVATORY

Where Is Unicode Used?
Asian Domains (cont.)

Unicode documents on web servers in Asian TLDs (2)					
TLD	Unicode docs (%)	Unicode domains (%)	TLD	Unicode docs (%)	Unicode domains (%)
mn	14.0	18.0	sg	21.3	20.4
mv	0.2	3.2	sy	6.1	9.1
my	14.4	14.4	th	4.0	13.2
np	48.7	14.7	tj	71.7	13.0
om	3.1	15.2	tm	48.5	8.1
pg	0.3	3.4	tp	3.4	9.4
ph	20.6	15.3	tr	5.6	9.4
ps	4.7	19.0	uz	3.0	3.7
qa	12.3	15.8	vn	69.1	74.5
sa	13.6	21.7	ye	0.9	10.6

UTF-8 Usage
by ccTLD in Asia

ITU-T

ccTLD	K pages	UTF-8
vn	1,556	70.55%
ir	516	57.14%
np	94	50.81%
bd	55	45.26%
lk	149	35.78%
mm	33	31.49%
ae	283	28.97%
in	1,463	24.04%
az	175	21.47%
ph	827	18.99%

ccTLD	K pages	UTF-8
sg	1,577	18.24%
pk	348	15.29%
mn	277	12.68%
my	1,932	11.66%
id	656	7.85%
il	15,197	5.82%
th	5,495	4.17%
kh	29	1.63%
la	198	0.09%
accessed Sep. - Oct. 2005		

LANGUAGE OBSERVATORY

source: Language-Observatory

End of presentation…

ITU-T

Terima kasih…..

Merci

Domo arigato gozaimasu

Syukran

THANK YOU…

"In the galaxy of languages, every word is a star."
… UNESCO

LANGUAGE OBSERVATORY