Unicode Spreading on the Web:

A case of Asian & African Domains

Pavol Zavarsky*, Wunna Ko Ko*
Yew Choong Chew*, Yoshiki Mikami*
and Tatsuo Kobayashi**

***Language Observatory**
c/o Nagaoka University of Technology
****Scholex Co., Ltd., Japan**

Internationalization and Unicode Conference 2006

CONTENTS

- Language Observatory
- Where the Unicode is used?
  - Case 1: Myanmar
  - Case 2: Bangladesh
  - Case 3: Vietnam
- Asian Languages Resource Network
- World Network for Linguistic Diversity

© Language Observatory 2006

Language Observatory
Objectives & Outputs

Objectives:

- To raise awareness on “Digital Language Divide” issues
- Support to the processing of the languages “falling trough the net”

Outputs:

- How many web pages are written by each language?
- What kind of charset is employed to encode it?
- How quickly UCS/Unicode is spreading?
- How a specific language community is linked with other language communities?

© Language Observatory 2006

Language Observatory
Members & Sponsors

Members:

- Nagaoka University of Technology, Japan
- Tokyo University of Foreign Studies, Japan
- Universiti Teknologi Malaysia, Malaysia
- Thai Computational Linguistics Laboratory, Thailand
- University of Miskolc, Hungary
- TDIL, Ministry of Information Technology, India
- Università degli Studi di Milano, Italy

Sponsors:

- Japan Science and Technology Agency
- UNESCO

© Language Observatory 2006

Language Observatory
System

- 40 Intel Xeon based PCs.
- 4GB memory.
- 250GB local disk space.
- 8TB crawler archive.

© Language Observatory 2006

Language Observatory
Crawler & Language Identifier

Crawler:

- UbiCrawler: Scalable, parallel, distributed crawler developed by our team members from Milano University

Language identifier:

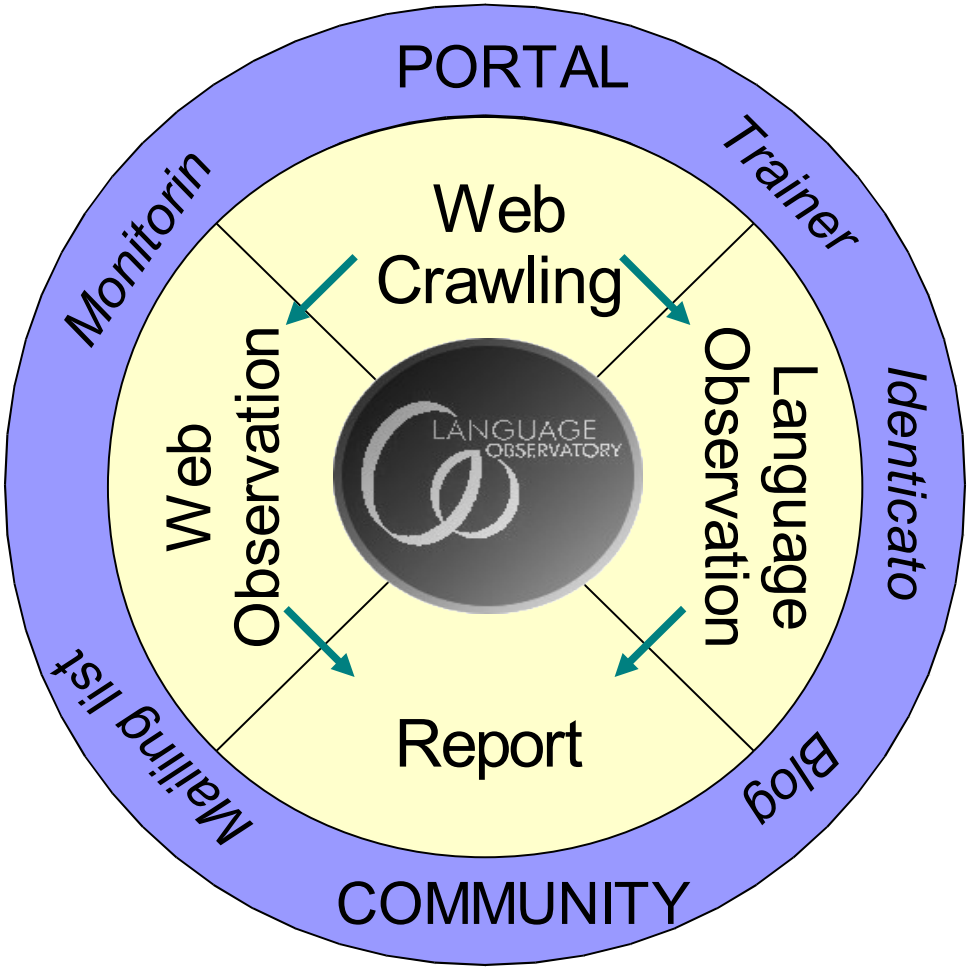
- Automatically identifies language/script/character encoding scheme of the page, > 330 LSE
- Training texts include “Universal Declaration of Human Rights” (new translations)

Web analysis:

- Web server, http status, web graph, web technology, etc.

© Language Observatory 2006

Language Observatory
World



© Language Observatory 2006

Where Unicode is used?
Asian & African summary

	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 2006

Where Unicode is used?
by ccTLD, Asian Domains

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

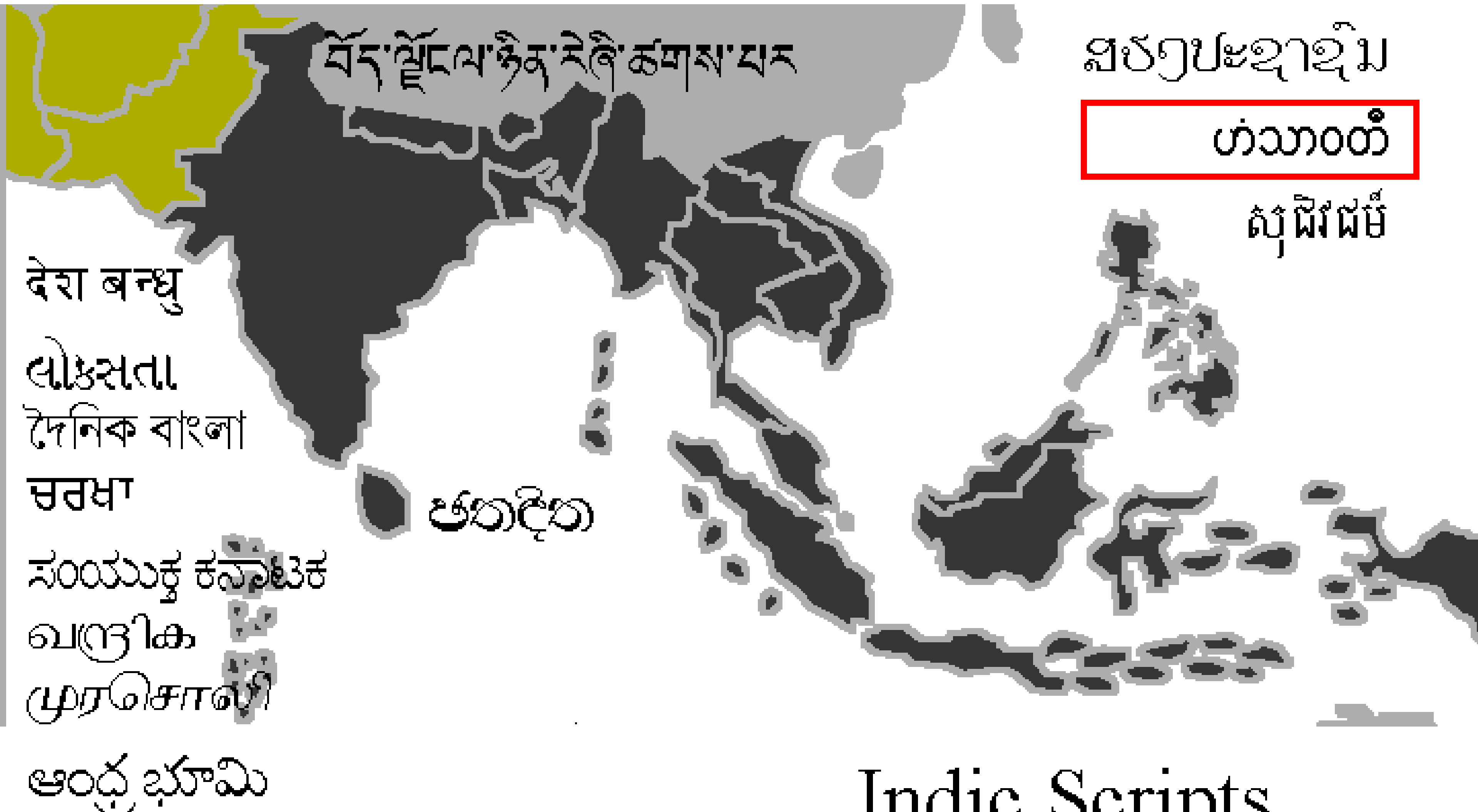
© Language Observatory 2006

Where Unicode is 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

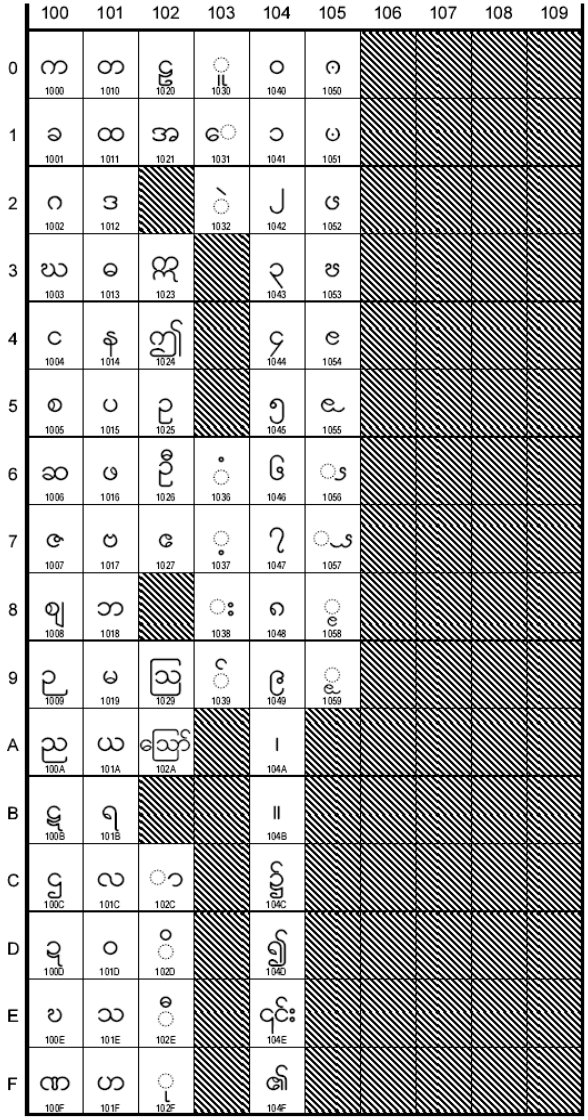
© Language Observatory 2006

Case 1: Myanmar
Myanmar Script



Indic Scripts

Unicode is here, but...



- Myanmar script has a position in Unicode since 1998
- Usage of Unicode is very low. Lowest among Southeast Asian Domains(0.2%)
- Most pages are written in English (even Government site).
- Some Myanmar Language texts are in ASCII format or Unicode area but not obeying Unicode rules.

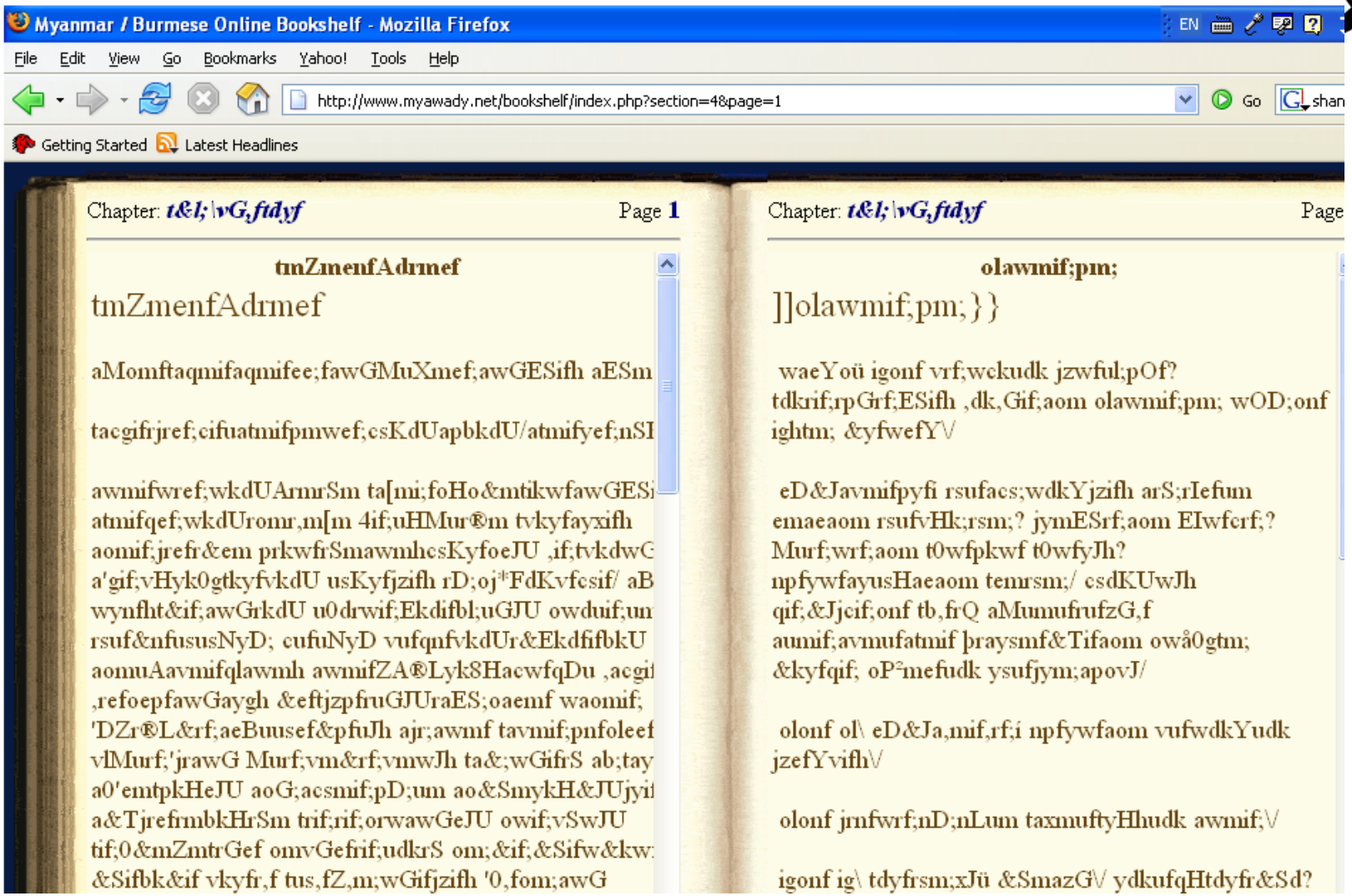
© Language Observatory 2006

Even .gov.mm site in English

http://www.myanmar.gov.mm

© Language Observatory 2006

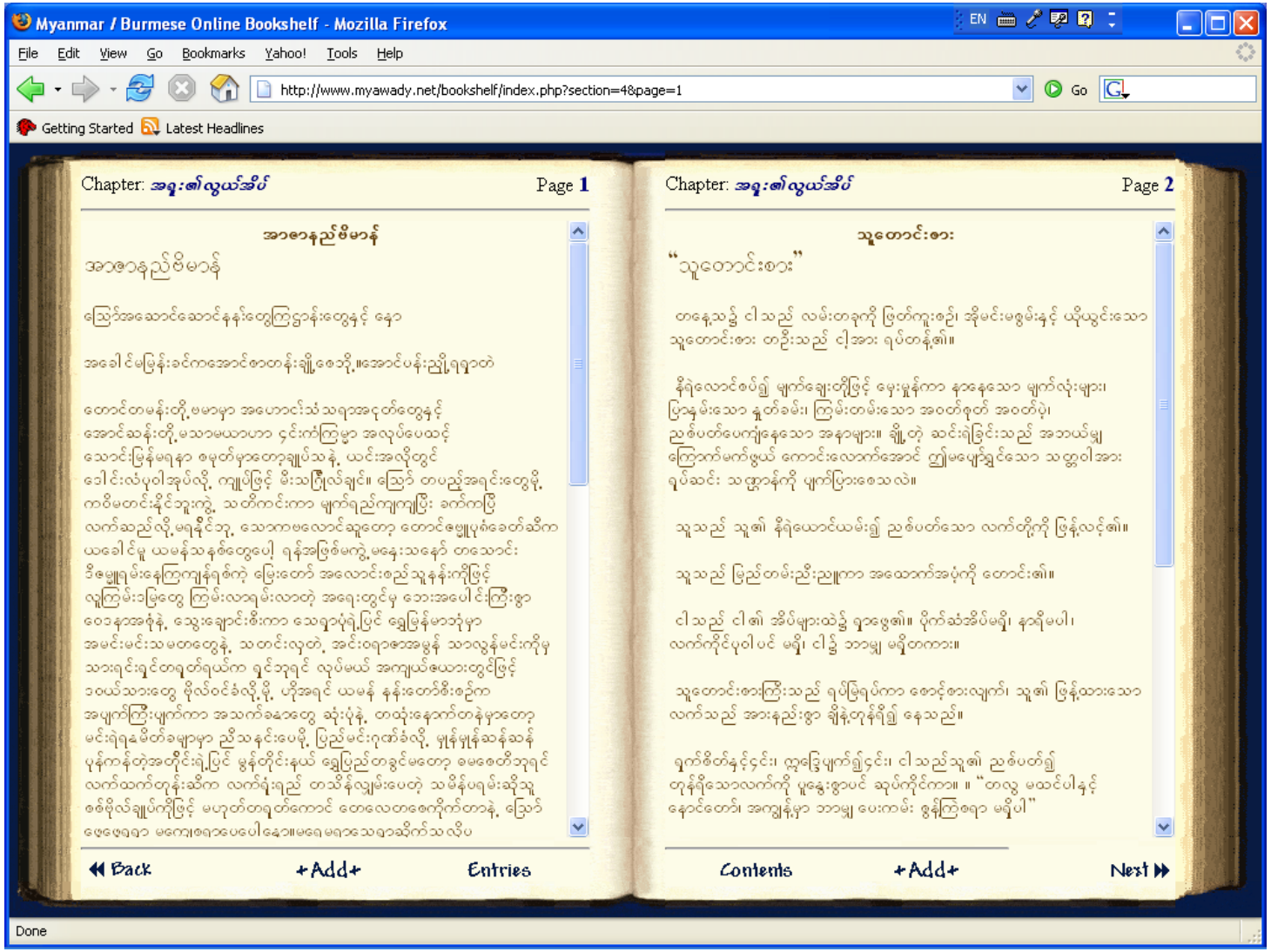
Various codes still in use



http://www.myawady.net/bookshelf/index.php?section=4&page=1

© Language Observatory 2006

Various codes still in use



http://www.myawady.net/bookshelf/index.php?section=4&page=1

© Language Observatory 2006

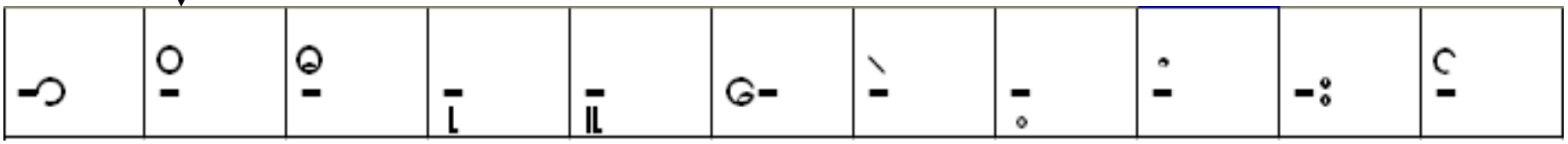
Burmese Script Set

Consonants

က	ခ	ဂ	ဃ	င
စ	ဆ	ဇ	ဈ	ည
ဋ	ဌ	ဍ	ဎ	ဏ
တ	ထ	ဒ	ဓ	န
ပ	ဖ	ဗ	ဘ	မ
ယ	ရ	လ	ဝ	သ
	ဟ	ဠ	အ	

Vowels

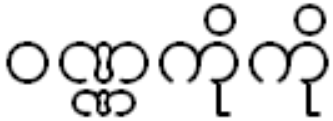
ၣ	ၤ	ၥ	ၦ	ၧ	ၨ	ၩ	ၪ	ၫ	ၬ	ၭ



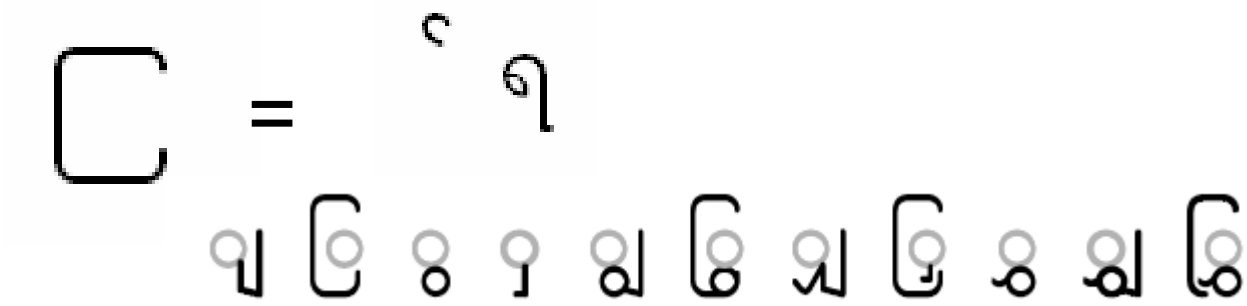
© Language Observatory 2006

Complexity of Burmese Script

Double Stacked



Ligatures



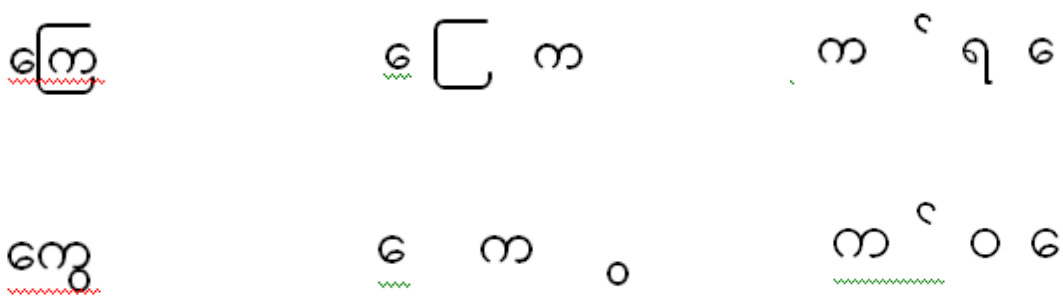
Glyph-shapes depends on consonants

© Language Observatory 2006

Burmese Script & Unicode

- Myanmar Scripts got position in Unicode in 1998. (U+1000 to U+109F) There is no standard encoding except Unicode.

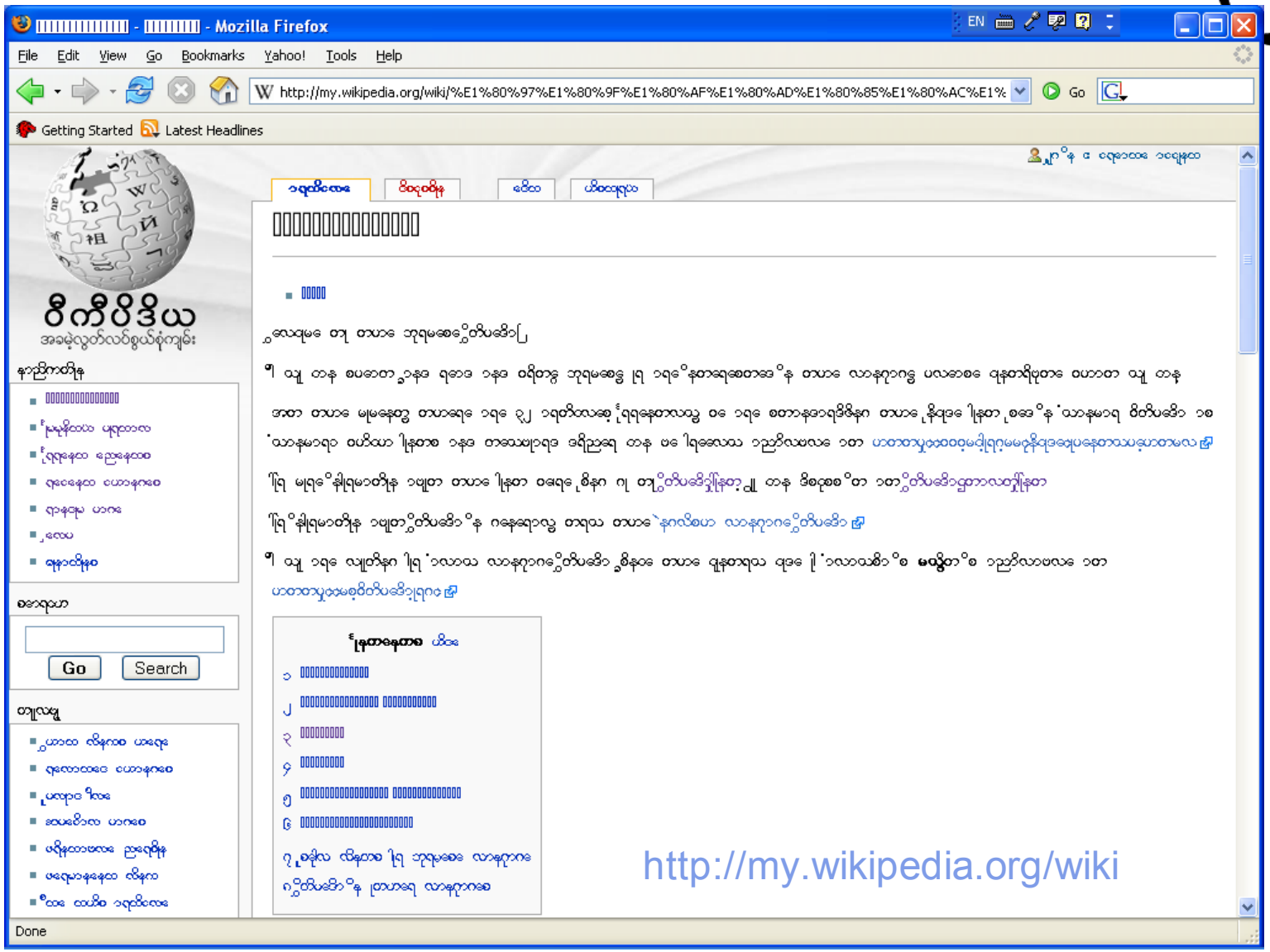
- Unicode encoding is phonetically ordered while other Myanmar TTFs, developed by replacing Glyph Position in ASCII, are graphically ordered.



Rendered | Graphical order | Phonetic order/ Storage

© Language Observatory 2006

Found only at Wikipedia



http://my.wikipedia.org/wiki

© Language Observatory 2006

Difficulties facing

Difficulty in migration into Unicode

- Many legacy Burmese texts in local encoding
- No conversion software, fully aware of Unicode rules
- No complete Unicode aware input system

Rendering System

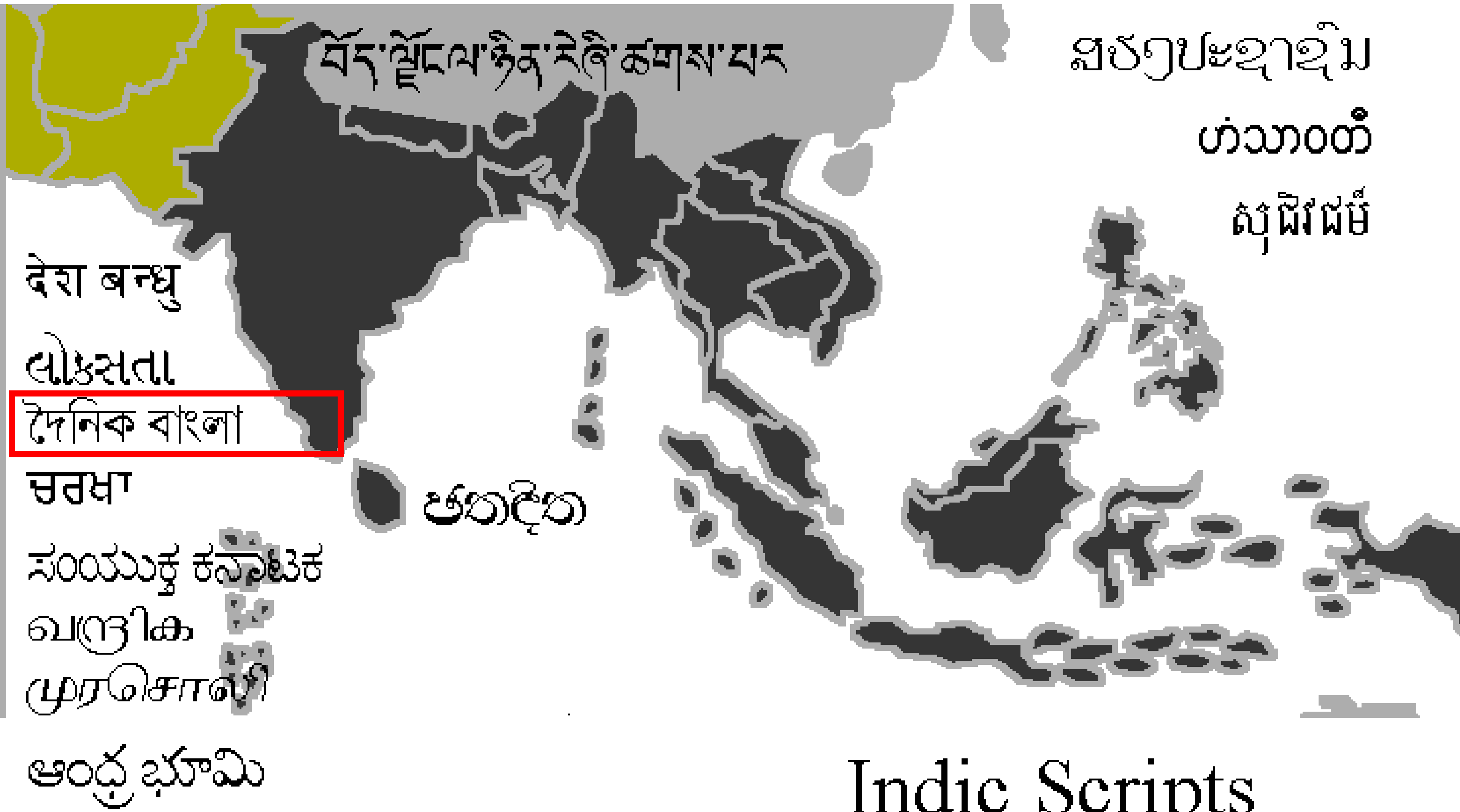
- M17N rendering system run on Linux O/S
- No complete rendering system yet on Windows O/S
- No web browser support complete rendering system

Availability of Fonts

- Win Innwa, Planet
- MPH Yangon, Myzedi

© Language Observatory 2006

Case 2: Bangladesh
Bengali Script



Mother language movement

“How can we forget the blood of my brother shed on 21st February”

© Language Observatory 2006

From Chaos to Unicode

Year	Standard and Policy
1945	Pakistan became independent (East Pakistan)
1952	"*Ekushay February*" (Bengali vs Urdu)
1971	Bangladesh became independent
1995	**BDS 1520: 1995** (8-bit national standard, graphically decomposed, glyph-like coded charset)
2000	**BDS 1520: 2000** (phonetic encoding-style totally new national standard)
	Open-source developers' pioneering efforts
2005	51% of web pages uses **Unicode**, but only 8% of domains.

© Language Observatory 2006

Case 3: Vietnam
Quôc ngu Alphabet

a	à	ả	ã	á	ạ
ă	ằ	ẳ	ẵ	ắ	ặ
â	ầ	ẩ	ẫ	ấ	ậ
e	è	ẻ	ẽ	é	ẹ
ê	ề	ể	ễ	ế	ệ
i	ì	ỉ	ĩ	í	ị
o	ò	ỏ	õ	ó	ọ
ô	ồ	ổ	ỗ	ố	ộ
ơ	ờ	ở	ỡ	ớ	ợ
u	ù	ủ	ũ	ú	ụ
ư	ừ	ử	ữ	ứ	ự
y	ỳ	ỷ	ỹ	ý	ỵ
đ					

- plus seven letters with diacritical mark
  ă, â, ê, ô, ơ, ư, đ
- five tone marks
  à, ả, ã, á, ạ
- total additions to ASCII
  = 134 (TCVN 5712)
  (pre-composed)
  = 52 (VNI, 6937-like)
  = 13 (USMARC)

© Language Observatory 2006

From Chaos to Unicode

Year	Standard and Policy
1969	**USMARC** (covers all combining diacritical marks used for Quôc ngu)
1987	**ISO 6937** (partially covers Quôc ngu diacritical marks, but lacks “hook”, “dot below”, Ơ and Ư)
	at least 30 encoding schemes observed VNI, VIQR, VPS, VISCII(RFC 1456), USMARC,...
1991	**TCVN 5712** (8-bit national standard)
2001	**TCVN 6909** (=ISO/IEC 10646)
	Government website migrate to TCVN 6909

© Language Observatory 2006

Where Unicode is used?
Asian Domains, 2004 to 2005

Trend in Unicode spreading in TLDs of Indochina and India (Example)				
	July 2004		Sept. 2005	
TLD	Unicode docs (%)	Unicode domains (%)	Unicode docs (%)	Unicode domains (%)
kh	1.4	4.1	1.7	5.6
mm	0.02	5.1	0.2	8.7
th	2.6	4.3	4.0	13.2
vn	63.2	57.1	69.1	74.5
in	6.0	5.6	31.2	24.6

© Language Observatory 2006

Asian Language Resource Network

Objectives:

- Promotion of internationalization of Internet-based activities
- Contribute to the creation of IT for under-resourced languages of developing parts of Asia
- Support of the development of technical conditions in order to facilitate the presence and use of the languages on the Internet
- Supported by:

UNESCO resolution: "The promotion and use of multilingualism and universal access to cyberspace"

© Language Observatory 2006

World Network for Linguistic Diversity

- Linguistic and cultural diversity in cyberspace identified as one of the priorities at the WSIS.
- A new institutional network created by the African Academy of Languages (ACALAN), under the auspices of the African Union
- The initiative started in the context of the WSIS; includes also organizations not solely focusing on cyber issues.
- Officially inaugurated by the Director General of UNESCO, H.E. Koichiro Matsuura
- Secretary General of the WNLD: Prof. Y. Mikami (Language Observatory)

© Language Observatory 2006

World Network for Linguistic Diversity
– Description and mission

- A multi-stakeholder network, involving civil societies, governments, international organizations, research centers and universities, media, the private sector and individuals
  - **African Academy of Languages (ACALAN)**
  - **Unicode IDN in Africa**
  - **Language Observatory**
  - **SIL International**
  - **Multilingual Internet Names Consortium (MINC)**
  - **International Telecommunication Union (ITU)**
  - **UNESCO**
  - **International Literacy Institute of the University of Pennsylvania**
  - **Union Africaine**

  …..

© Language Observatory 2006

World Network for Linguistic Diversity - Objectives

- Encourage governments and institutions to adopt and implement measures enhancing equitable multilingualism;
- Promote mother tongue-based bilingual and/or multilingual education at all levels of education everywhere as a means of promoting both social and gender equality;
- Promote software localization and equal access of all languages to cyberspace;
- Facilitate the empowerment of language communities worldwide in developing and using their own languages;
- Contribute to the creation and sharing of language resources;
- Observe the implementation of language policies and serve as a focal point for linguistic research projects.

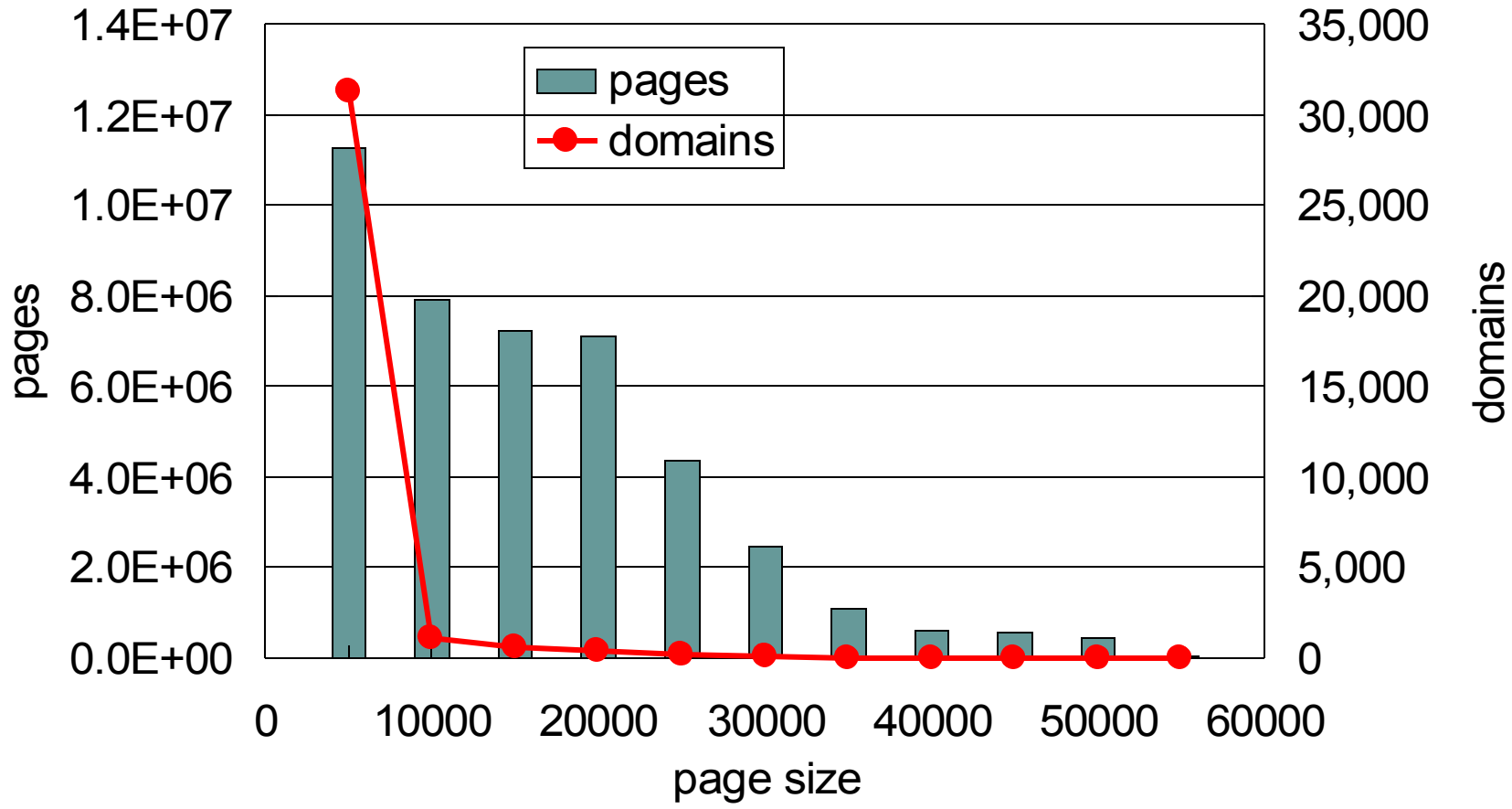
© Language Observatory 2006

Where Unicode is used?
by ccTLD, African Domains

Unicode documents on web servers in African TLDs					
TLD	Unicode docs (%)	Unicode domains (%)	TLD	Unicode docs (%)	Unicode domains (%)
ac	3.1	1.6	ml	6.0	5.6
bf	0.4	3.2	mr	5.8	2.8
bw	2.1	4.1	mu	6.3	13.6
ci	17.4	58.5	na	14.1	8.2
cv	0.6	5.7	ng	16.3	12.3
dz	0.7	6.7	rw	11.2	6.1
eg	3.2	11.5	sc	23.5	23.4
gh	0.7	3.9	sh	23.7	2.2
gm	17.0	18.5	sn	0.3	2.8
ke	9.4	8.4	st	29.7	5.2
ly	9.5	1.9	sz	1.2	6.9
ma	4.0	8.5	tn	11.5	7.8
mg	0.1	5.1	za	10.6	5.1

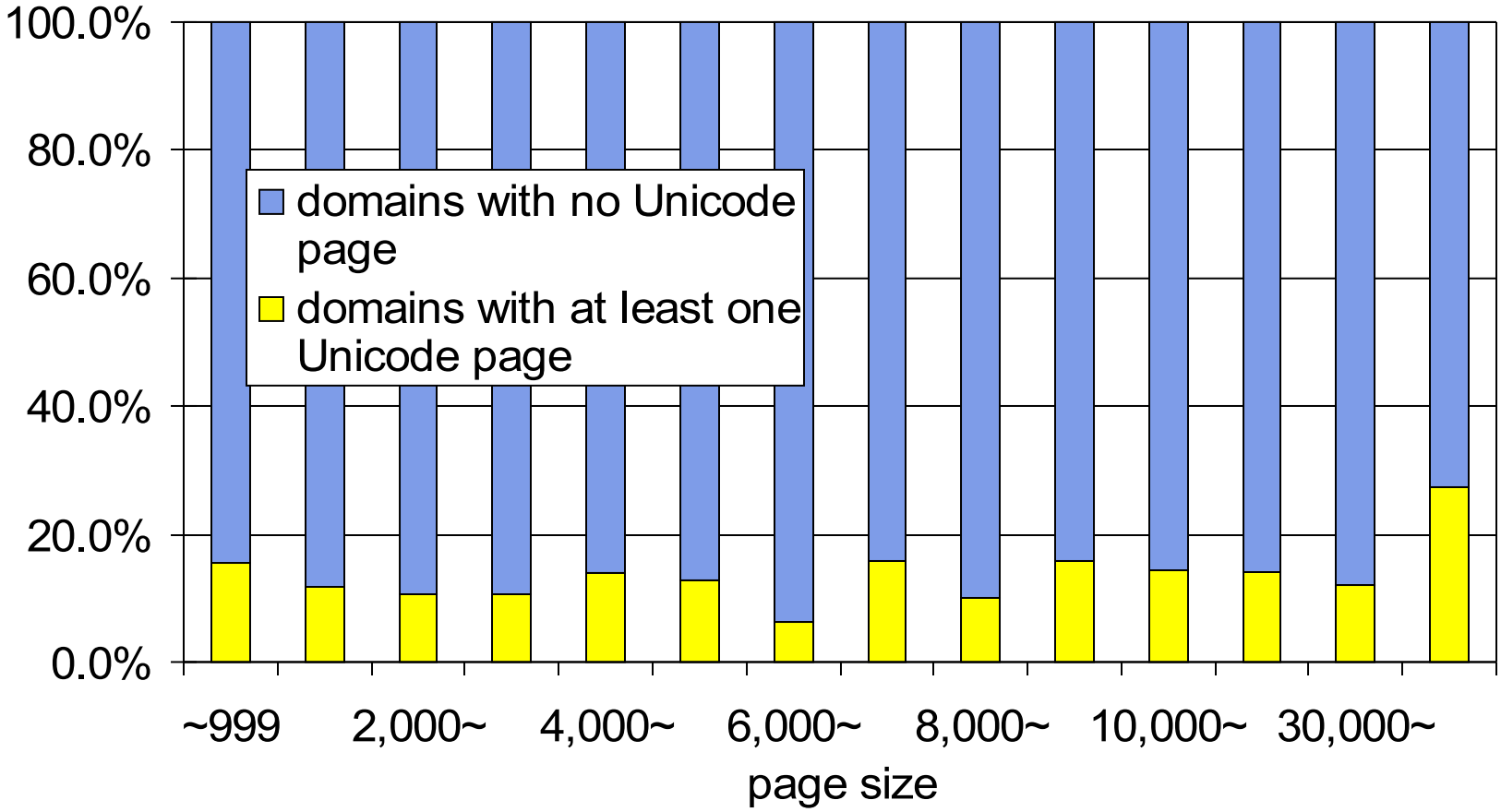
© Language Observatory 2006

Page-size distribution of Web



© Language Observatory 2006

Who Uses Unicode?
by domain page-size



© Language Observatory 2006

CONCLUSION

- 10 - 15% web pages use Unicode in Asian and African domains.
- Usage rate varies depending on country, from almost nothing to as high as 70%.
- Driving force:
  - **Government initiatives: Vietnam case**
  - **OSS developers, portal sites, etc.**
- Many countries are still in need of appropriate localization tools.

© Language Observatory 2006

Acknowledgement

- The study was made possible by the financial support of the Japan Science and Technology Agency (JST) under the RISTEX program.

© Language Observatory 2006