Digital Language Divide

Measuring Linguistic Diversity on the Internet

UNESCO/UNU Conference on Globalization and Languages: Building on our Rich Heritage

Tokyo, Japan, 27 – 28 August 2008

Yoshiki Mikami

Leader, Language Observatory Project

Executive Committee member of MAAYA

Professor, Nagaoka Univ. of Technology

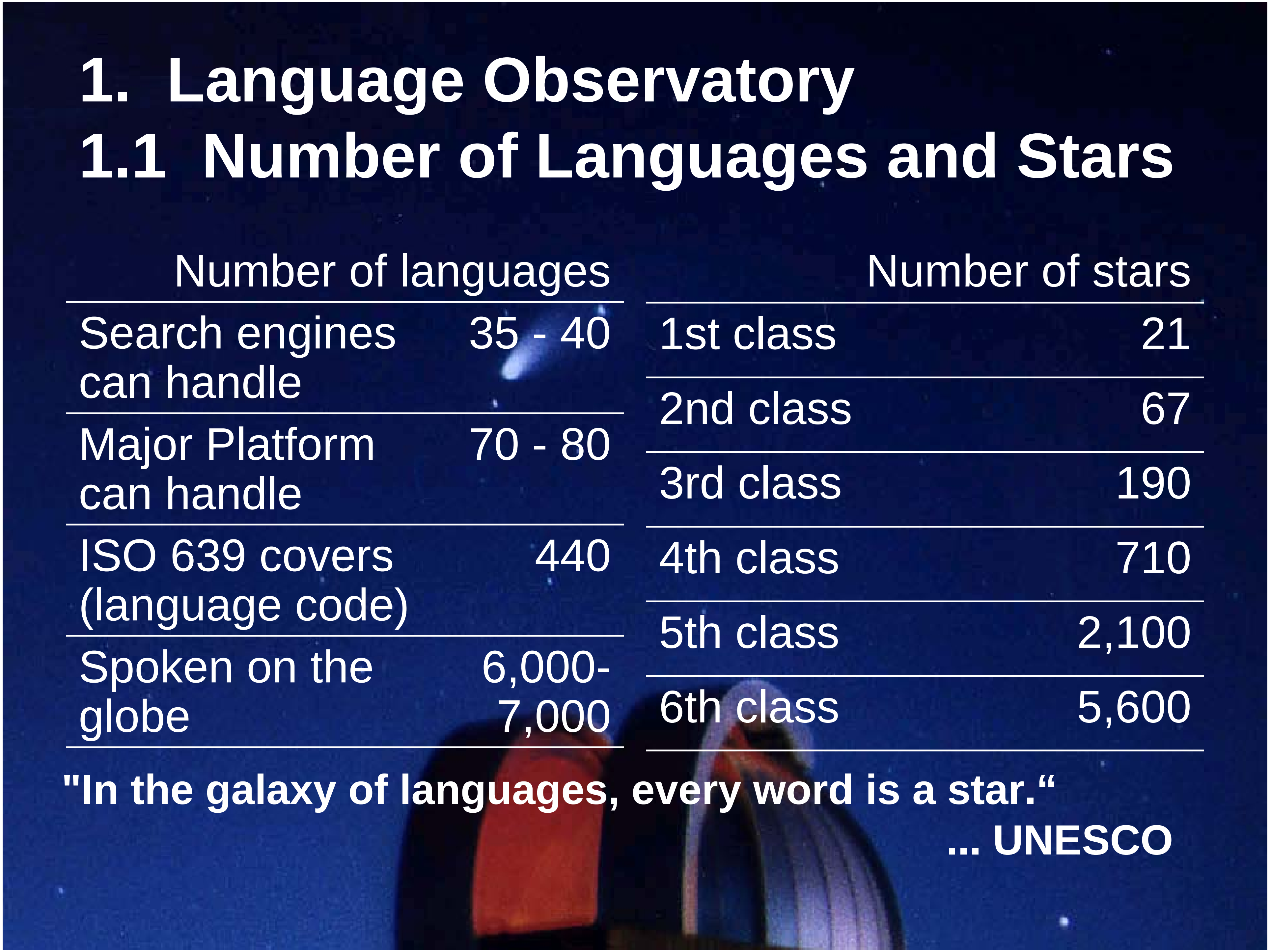
LANGUAGE OBSERVATORY

maaya

world network
for linguistic
diversity

Outlines

1. **Language Observatory**
   - **1.1 Language and Stars**
   - **1.2 How It Functions?**
2. **Survey Snapshots**
   - **2.1 Asia**
   - **2.2 Africa**
3. **Factors Behind**
   - **3.1 Economic factor**
   - **3.2 Technical factor**
   - **3.3 Socio-cultural factor**
4. **Conclusion**



1. Language Observatory

1.1 Number of Languages and Stars

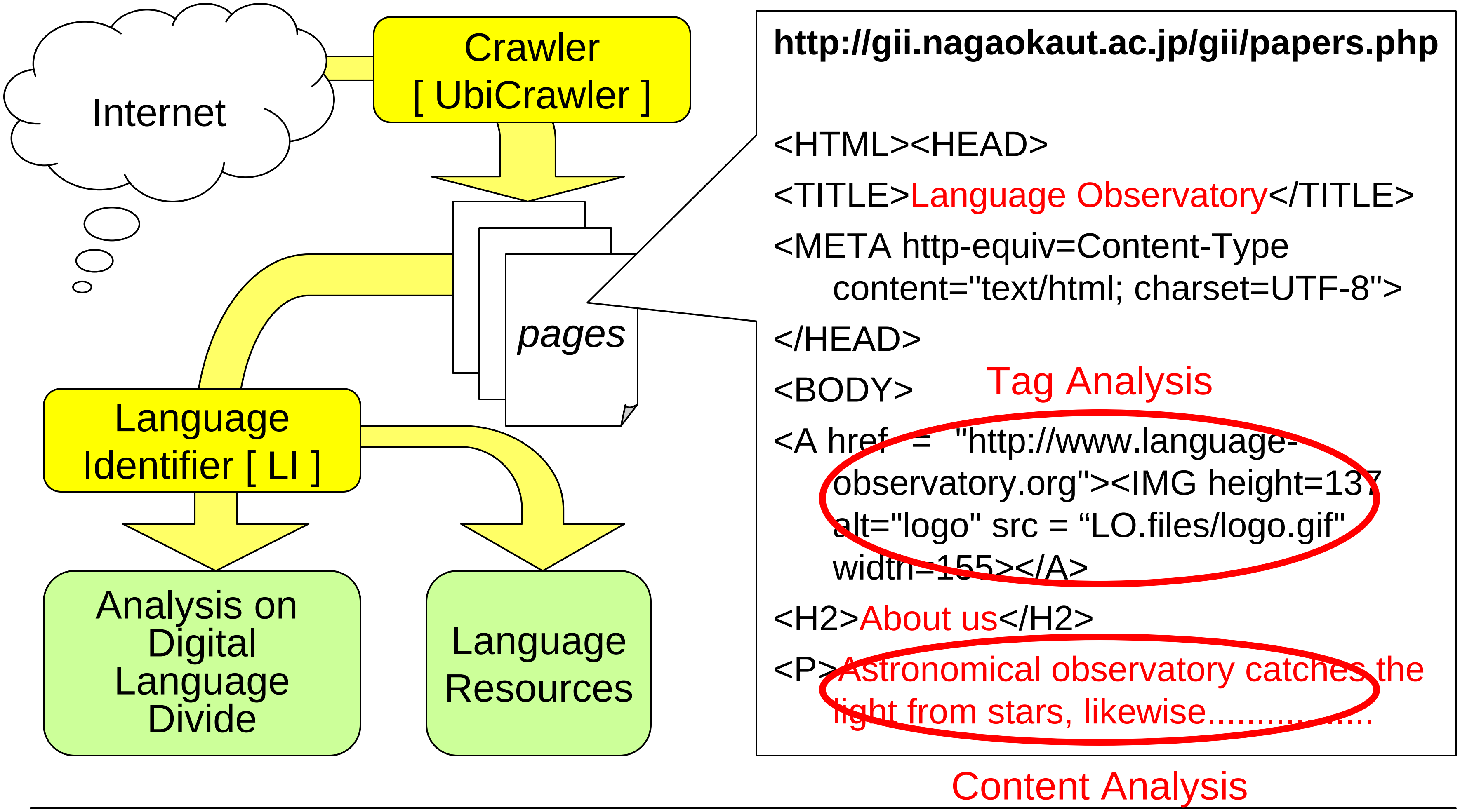
	Number of languages
Search engines can handle	35 - 40
Major Platform can handle	70 - 80
ISO 639 covers (language code)	440
Spoken on the globe	6,000-7,000

	Number of stars
1st class	21
2nd class	67
3rd class	190
4th class	710
5th class	2,100
6th class	5,600

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

... UNESCO

1. Language Observatory
1.2 How It Functions?



http://gii.nagaokaut.ac.jp/gii/papers.php

```
<HTML><HEAD>
<TITLE>Language Observatory</TITLE>
<META http-equiv=Content-Type
     content="text/html; charset=UTF-8">
</HEAD>
<BODY>
<A href  =  "http://www.language-
     observatory.org"><IMG height=137
     alt="logo" src = “LO.files/logo.gif"
     width=155></A>
<H2>About us</H2>
<P>Astronomical observatory catches the
     light from stars, likewise...............
```

Tag Analysis

Content Analysis

Unit of Identification
Language+Script+Encoding

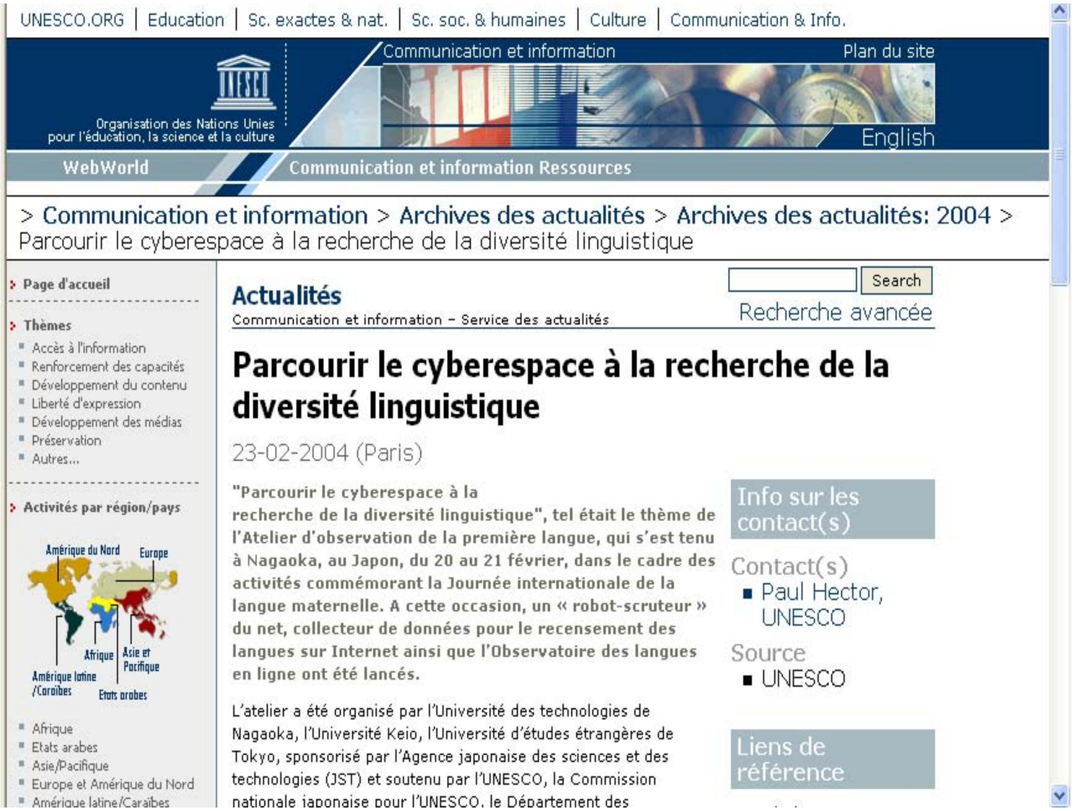
Language	Script	Encoding
Dari	Arabic	UTF-8
Farsi	Arabic	UTF-8
Hindi	Devanagari	UTF-8
Hindi	Devanagari	Arjun
Hindi	Devanagari	Shusha
Hindi	Devanagari	Shivaji
Azeri	Latin	Latin-1
Azeri	Cyrillic	KOИ-R
Azeri	Arabic	ASMO

Difference of language

Differnce of Encoding

Difference of Script

The Project Launched in 2004 on Int’l Mother Language Day



UNESCO.ORG | Education | Sc. exactes & nat. | Sc. soc. & humaines | Culture | Communication & Info.

Organisation des Nations Unies pour l'éducation, la science et la culture

Communication et information — Plan du site — English

WebWorld — Communication et information Ressources

> Communication et information > Archives des actualités > Archives des actualités: 2004 > Parcourir le cyberespace à la recherche de la diversité linguistique

- Page d'accueil
- Thèmes
  - Accès à l'information
  - Renforcement des capacités
  - Développement du contenu
  - Liberté d'expression
  - Développement des médias
  - Préservation
  - Autres...
- Activités par région/pays

- Afrique
- Etats arabes
- Asie/Pacifique
- Europe et Amérique du Nord
- Amérique latine/Caraïbes

Search — Recherche avancée

Actualités

Communication et information - Service des actualités

Parcourir le cyberespace à la recherche de la diversité linguistique

23-02-2004 (Paris)

"Parcourir le cyberespace à la recherche de la diversité linguistique", tel était le thème de l'Atelier d'observation de la première langue, qui s'est tenu à Nagaoka, au Japon, du 20 au 21 février, dans le cadre des activités commémorant la Journée internationale de la langue maternelle. A cette occasion, un « robot-scruteur » du net, collecteur de données pour le recensement des langues sur Internet ainsi que l'Observatoire des langues en ligne ont été lancés.

L'atelier a été organisé par l'Université des technologies de Nagaoka, l'Université Keio, l'Université d'études étrangères de Tokyo, sponsorisé par l'Agence japonaise des sciences et des technologies (JST) et soutenu par l'UNESCO, la Commission nationale japonaise pour l'UNESCO, le Département des

Info sur les contact(s)

Contact(s)
- Paul Hector, UNESCO

Source
- UNESCO

Liens de référence

UNESCO reported the launch of the project

UNESCO Recommendation

Recommendation concerning the Promotion and Use of Multilingualism and Universal Access to Cyberspace, October 2003

[PREAMBLE]

- Noting that linguistic diversity in the global information networks and universal access to information in cyberspace are at the core of contemporary debates and can be a determining factor in the development of a knowledge-based society,

Milestones, 2003 to 2007

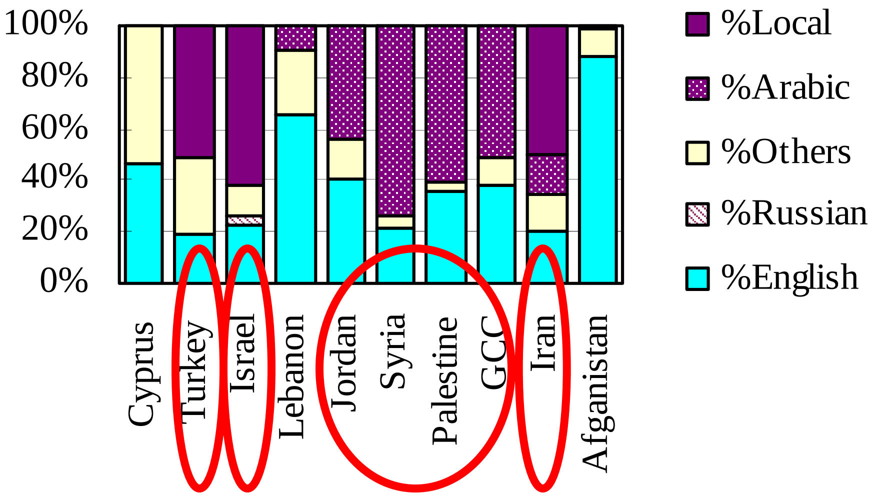
Oct. 2003	**UNESCO** Adopted “Cyberspace Recommendation”
Oct. 2003	Project started by the support of Japan Science and Technology Agency (JST)
Feb. 2004	The First Language Observatory Workshop
Jun. 2004	Started to collect web data by **“UbiCrawler”**
Aug. 2005	The First version of **”Language Identification Module”**
Nov. 2005	WSIS Tunis meeting
Feb. 2006	World Network for Linguistic Diversity **(MAAYA)** created
Jun. 2006	Workshop at Bamako, Mali on African Survey
Feb. 2007	Workshop at UNESCO, Paris
Sep. 2007	JST Funded Project Completed

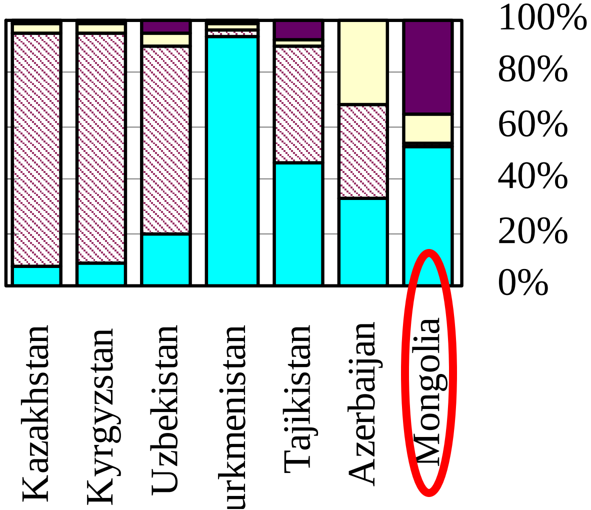
Expert Collaboration
Case of African Survey

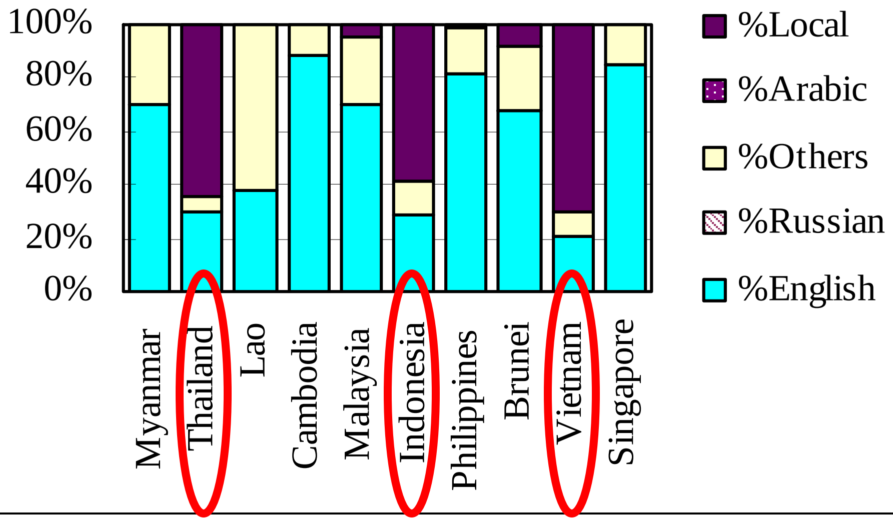
June 26-28, 2006 at Bamako, Mali

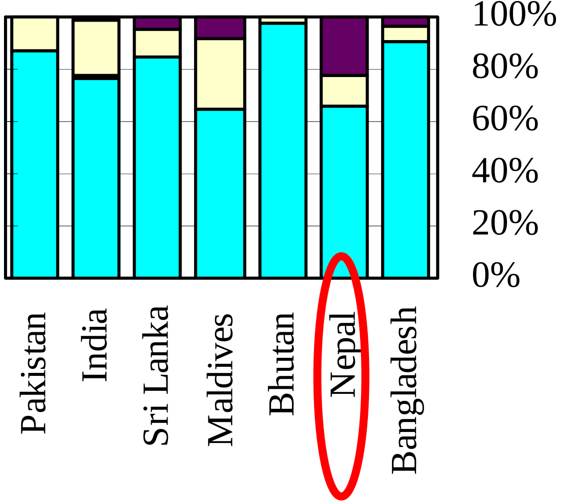
ACALAN
Mali
Algeria
Burkina Faso
Ethiopia
Kenya
Malawi
Nigeria
Tunisia
CNRS, France

2 Survey Snapshots
2.1 Asia









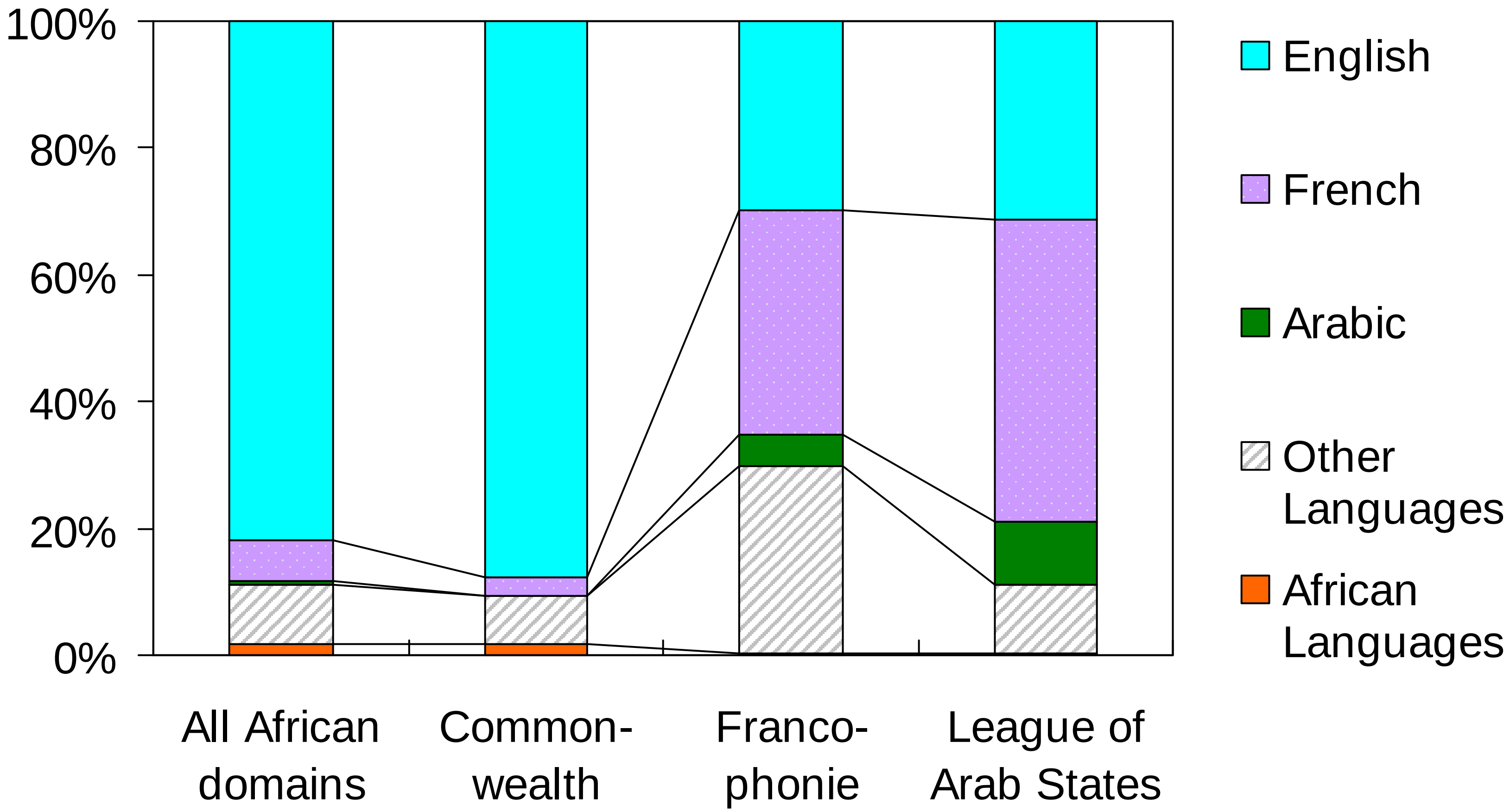
Estimated number of pages Top 10 Asian languages

Language	Script	Speaker population	pages
Hebrew	Hebrew	4,612,000	11,957,314
Thai	Thai	21,000,000	7,752,785
Turkish	Latin	59,000,000	3,959,328
Vietnamese	Latin	66,897,000	2,006,469
Arabic	Arabic	280,000,000	1,671,122
Tatar	Latin	7,000,000	1,575,442
Farsi	Latin	33,000,000	1,293,880
Javanese	Latin	75,000,000	1,267,981
Indonesian	Latin	140,000,000	866,238
Malay	Latin	17,600,000	432,784

Note: Chinese, Korean & Japanese domains are excluded. As of October 2006

as of June 2006

2 Survey Snapshots
2.1 Africa

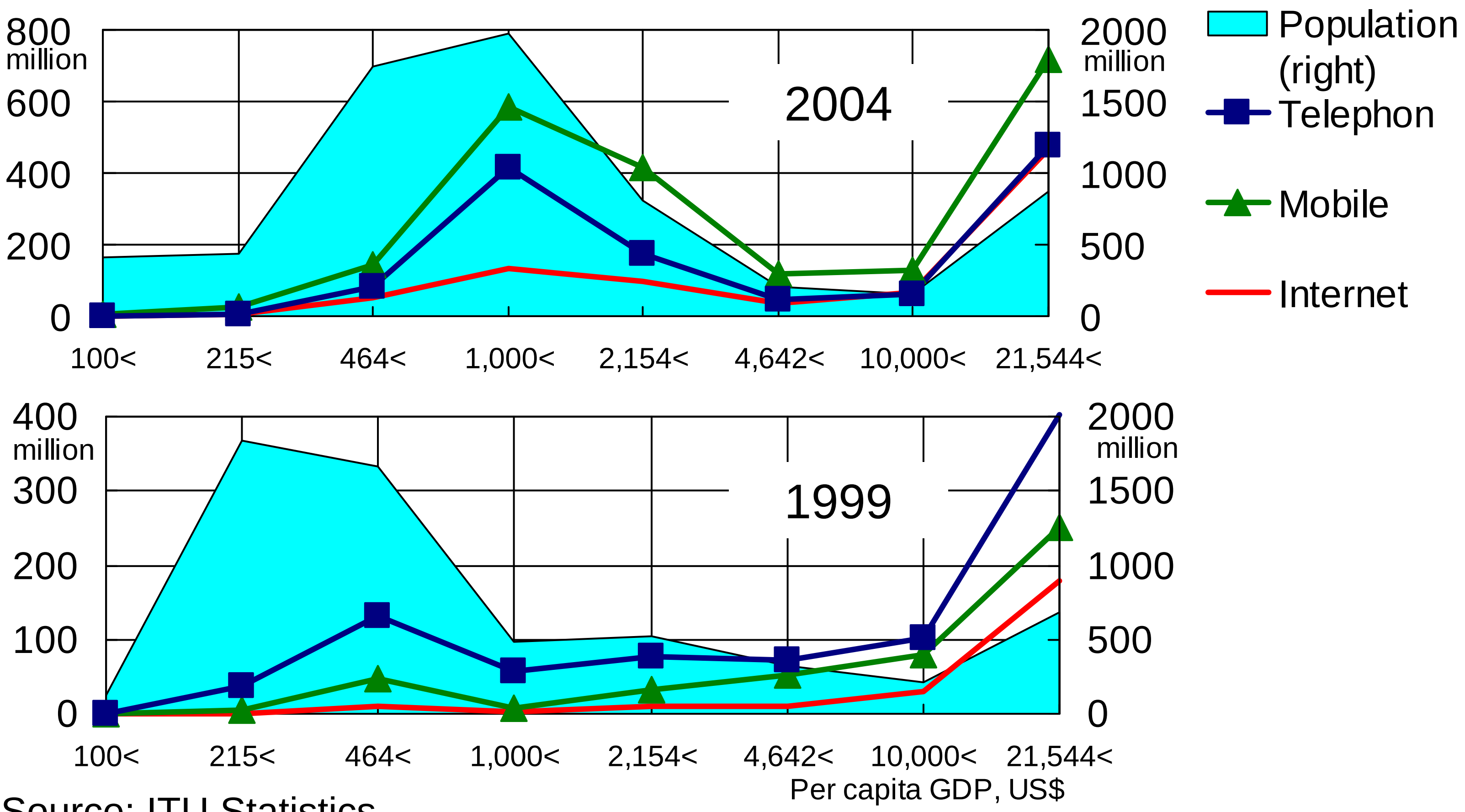


Estimated number of pages Top 10 African languages

language	script	speaking region	pages
Malagasy	Latin	Madagascar	5,382
Swahili	Latin	Tanzania	5,170
Afrikaans	Latin	South Africa, Namibia	1,775
Krio	Latin	Gambia, Sierra Leone	1,575
Kinyarwanda	Latin	Rwanda	1,059
Shona	Latin	Zimbabwe, Mozambique	538
Somali	Latin	Somalia	396
Siswati	Latin	Swaziland	335
Oshiwanbo	Latin	Namibia, Angola	264
Rundi	Latin	Burundi	252

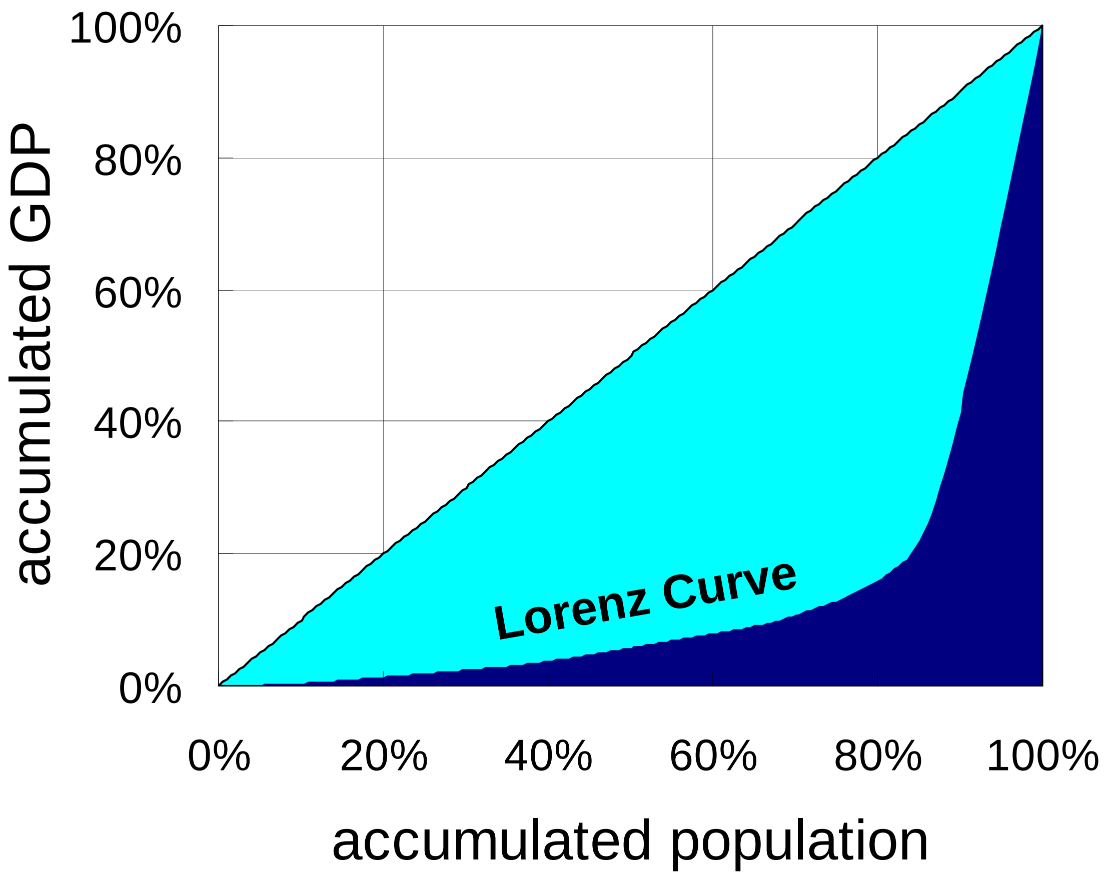
Note: South Africa is excluded. As of October 2006

3. How to Interpret it?
3.1 Economic Context



Source: ITU Statistics

3. Factors Behind
3.1 Economic Factor



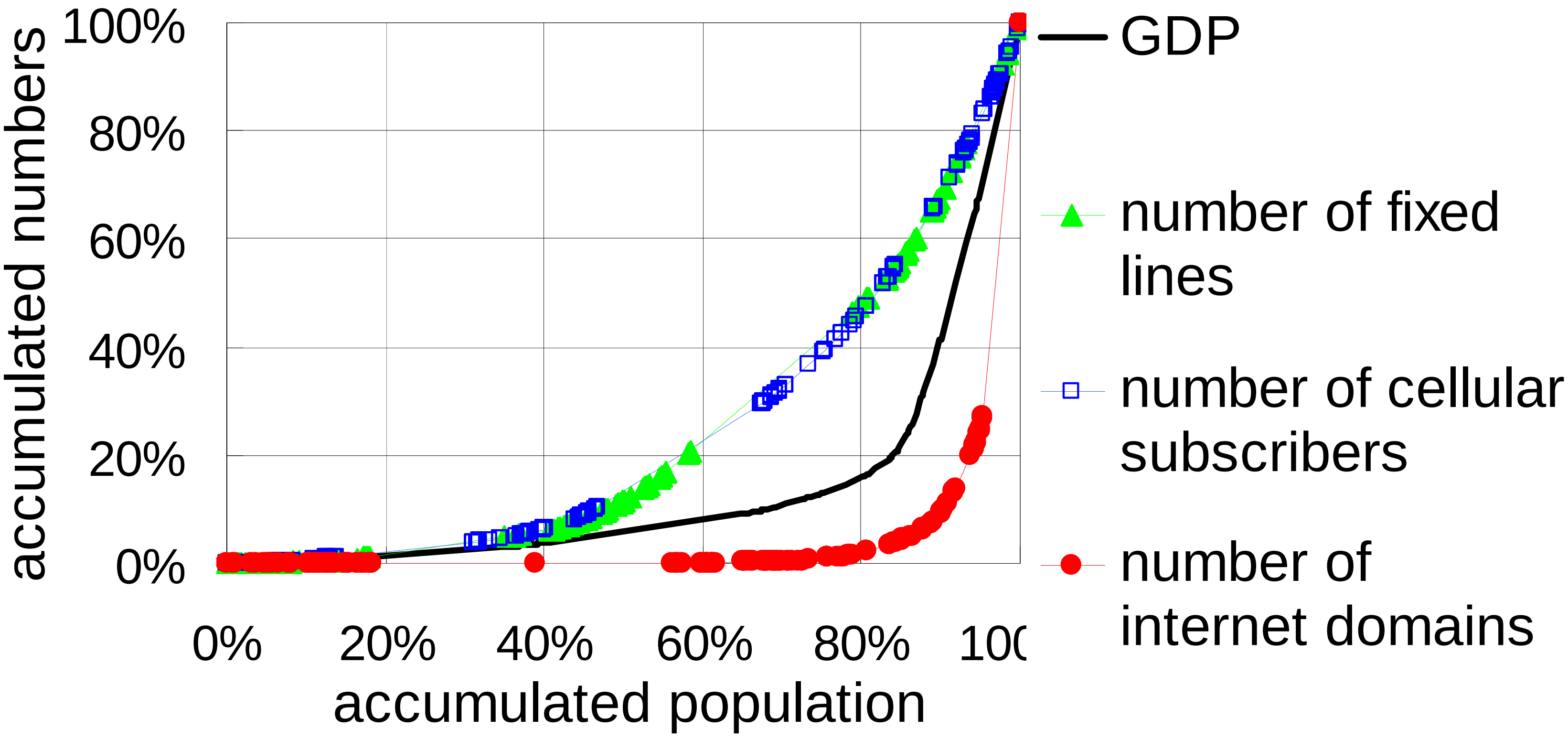
Gini Coefficient

$$= \frac{\square}{\square + \blacksquare}$$

0: perfect equality

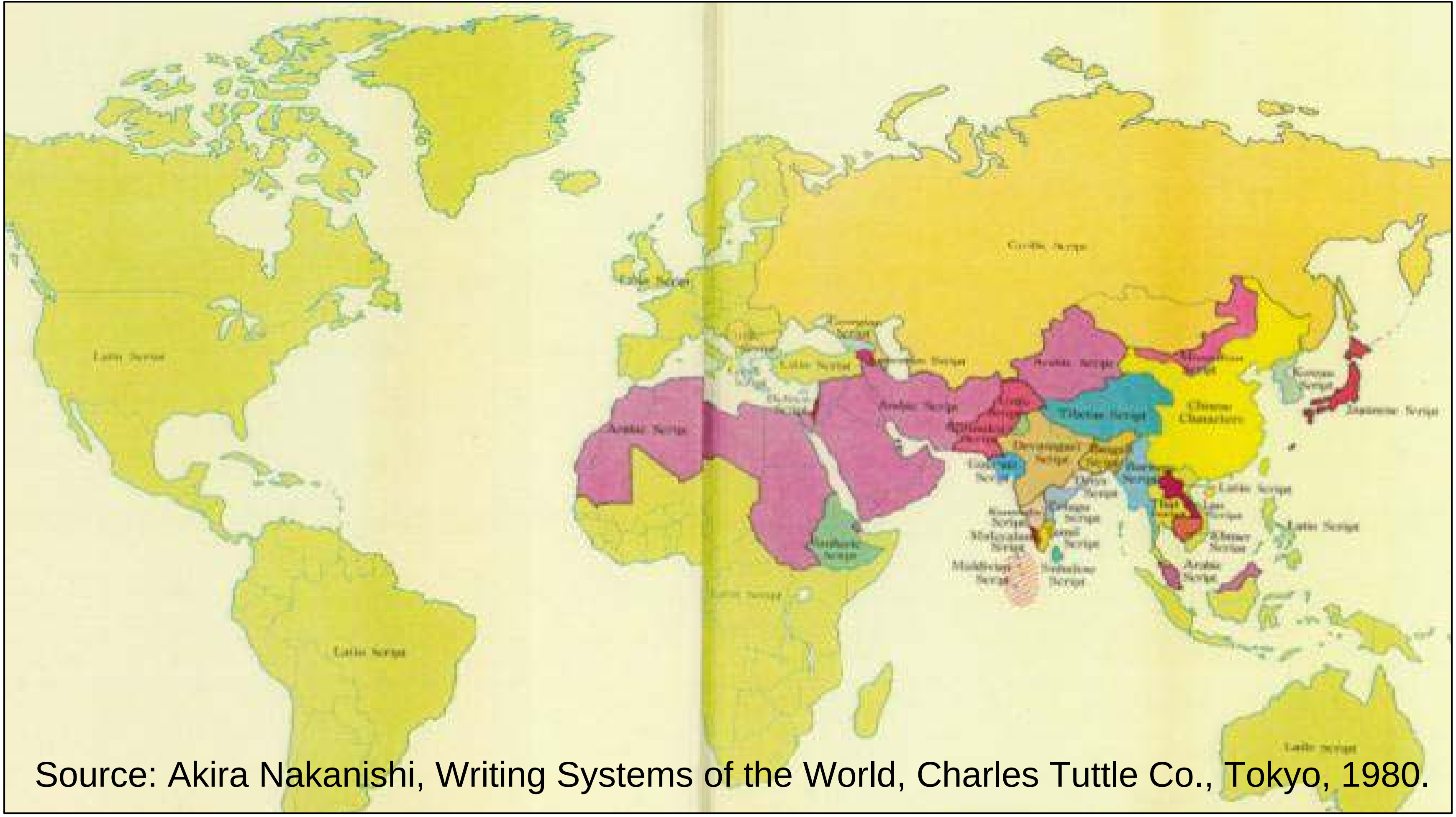
1: perfect inequality

Telephony has been improved, but Internet is…



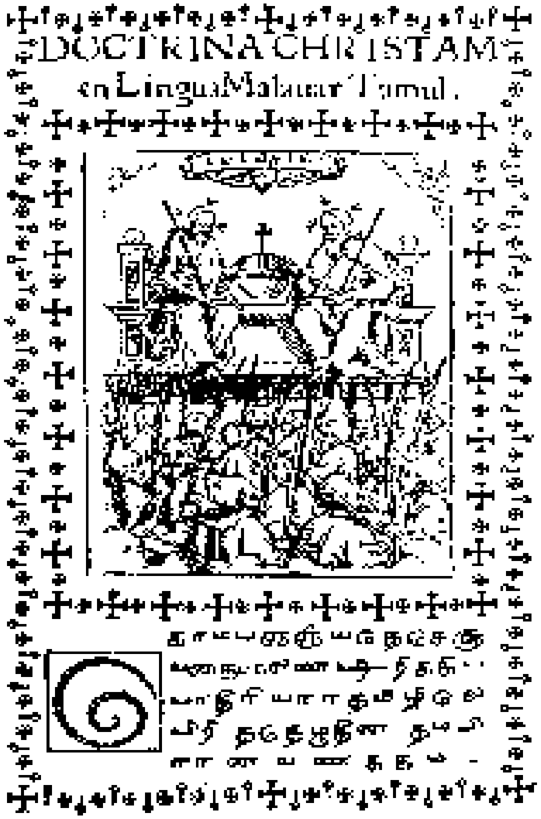
Gini-coefficient: Telephony 0.51 < GDP 0.73 < Internet 0.91

3.2 Technical Factor
World Map of Scripts



Source: Akira Nakanishi, Writing Systems of the World, Charles Tuttle Co., Tokyo, 1980.

A Jesuit Friar’s letter, 1608
Six hundred versus 24



DOCTRINA CHRISTAM

en Lingua Malauar Tamul.

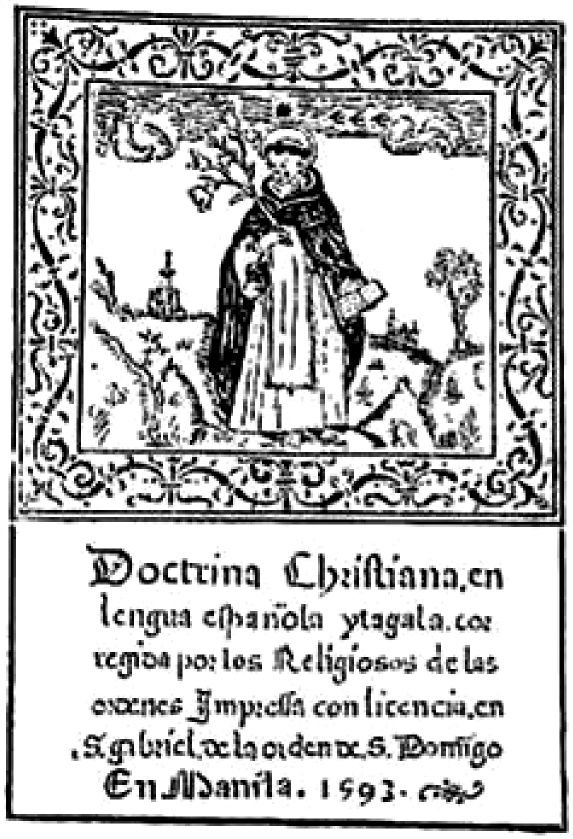
[illegible]

Doctrina Christam
in Tamil, 1578

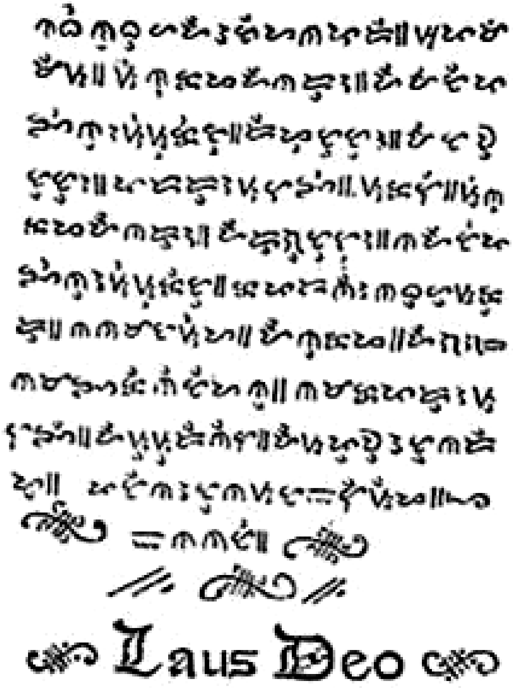
"Before I end this letter I wish to bring before Your Paternity's mind the fact that for many years I very strongly desired to see in this Province some books printed in the language and alphabet of the land, as there are in Malabar with great benefit for that Christian community. And this could not be achieved for two reasons; the first because it looked impossible to cast so many moulds amounting to six hundred, whilst as our twenty-four in Europe."

source: Priolkar, The Printing Press in India, Bombay, 1958

Case of Tagalog: The script was finally lost



Doctrina Chriſtiana,en
lengua eſpañola ytagala.cor
regida por los Religioſos de las
ordenes Impreſſa con licencia,en
S.gabriel,dela orden de S.Domingo
En Manila. 1593.

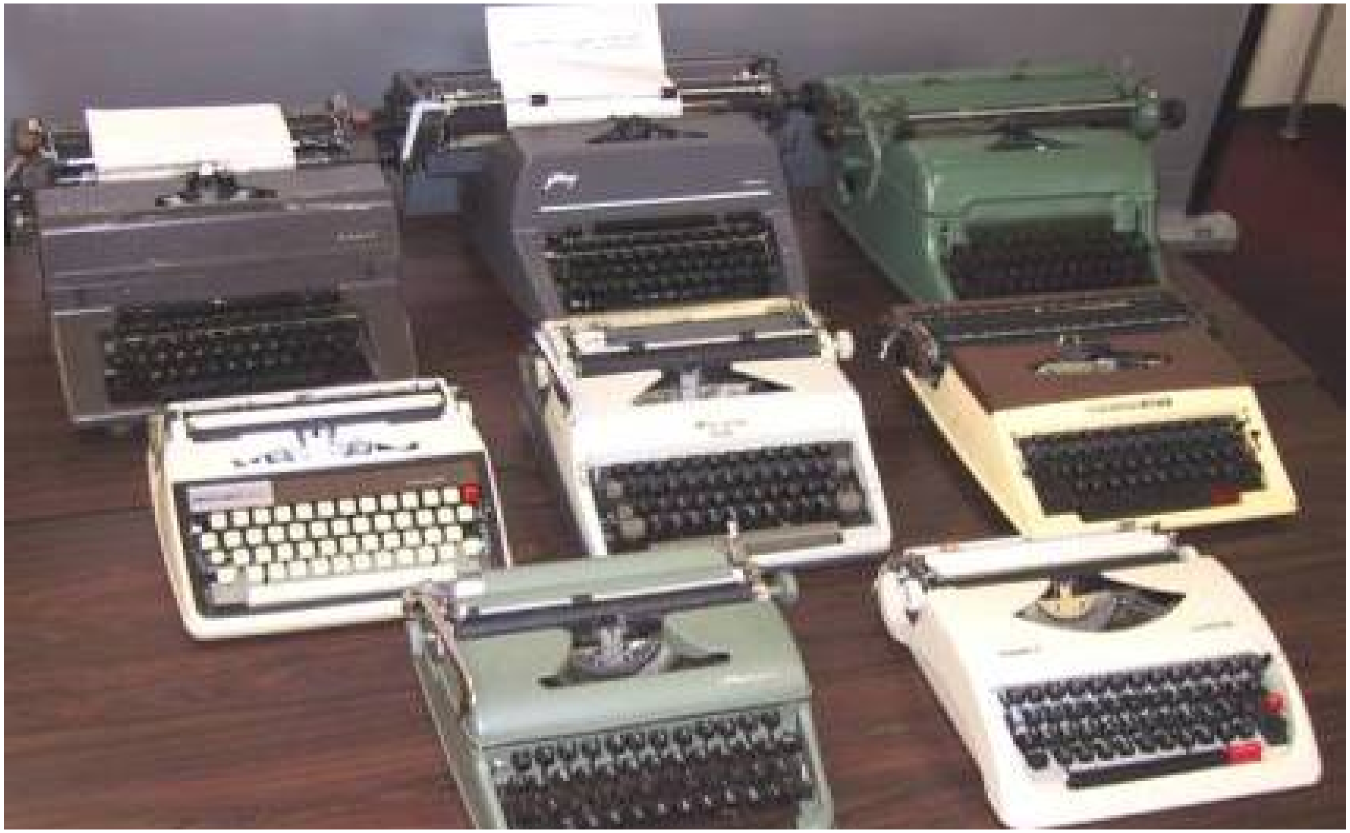


Laus Deo

Philippines
postal stamp
issued in 1995

“Doctrina Christiana”, bi-lingual version, printed in Tagalog by Tagalog script / in Tagalog by Latin script / in Spanish by Latin script. (1593)

Asian Language Typewriter Collection



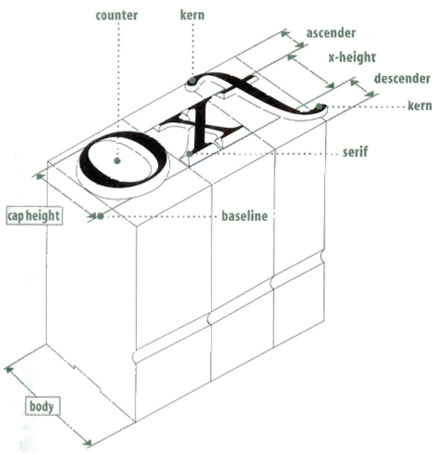
top to bottom

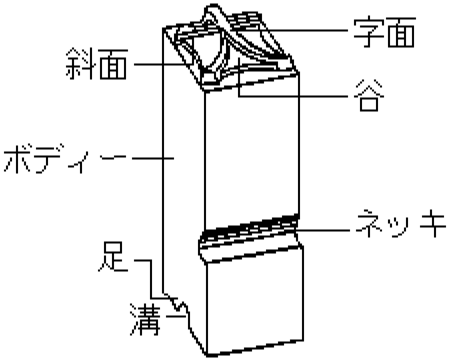
Tamil, Bengali, Sinhalese /

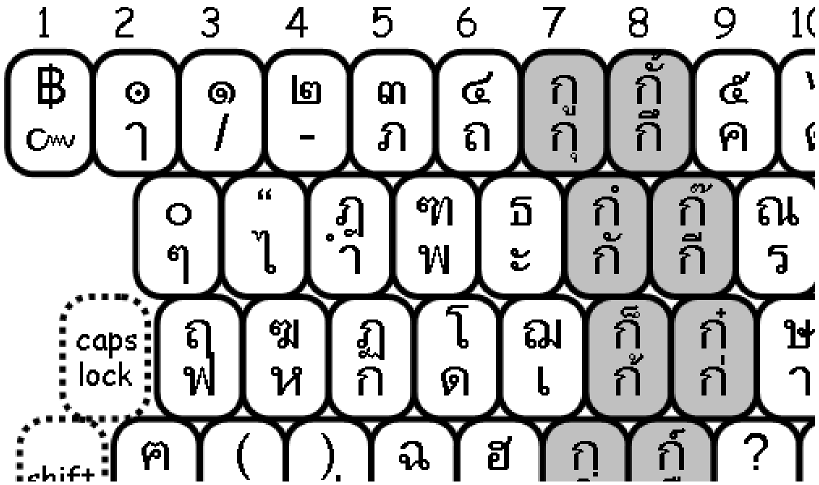
English, Hindi, Korean /

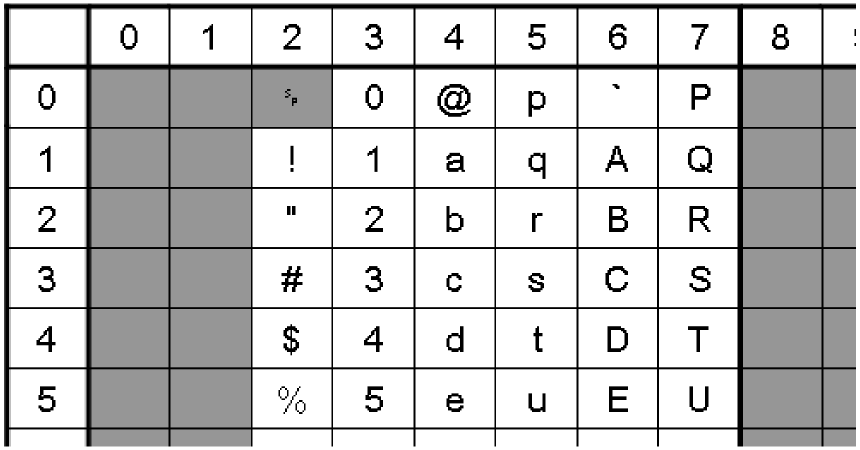
Myanmar, Thai

Localization Problem









	0	1	2	3	4	5	6	7	8
0			SP	0	@	p	`	P	
1			!	1	a	q	A	Q	
2			"	2	b	r	B	R	
3			#	3	c	s	C	S	
4			$	4	d	t	D	T	
5			%	5	e	u	E	U	

- *“Language Localization”* has been the key obstacle to the use of new information technologies since type printing age.

Encoding Chaos leads to delay of localization

Language	Standard encoding and its share	Examples of other encodings found [footnote]
Turkish	ISO 8859 (99.5%)	
Hebrew	ISO 8859 (87.7%)	
Vietnamese	UTF-8 (96.4%)	TCVN, VIQR, VPS
Thai	TIS 620 (97.3%)	
Mongolian	UTF-8 (95.5%)	Latin-Cyrillic
Sinhala	UTF-8 (44.5%)	Metta, Kaputa, etc.
Telugu	UTF-8 (16.6%)	Shree, TLH, etc.
Tamil	UTF-8 (14.9%)	Amudham, Kumudam, Shree, Vikatan, etc.
Burmese	UTF-8 (0.7%)	WinResearcher, etc.

note: Local proprietary encodings are shown in this table by names of font (families). as of June 2006

3.3 Socio-cultural Factor
Four Domains of languages

Personal domain	Public domain	Occupational domain	Educational domain
Conversation, mail, phone, blog, magazines, newspaper, novel, songs, etc.	Official documents, laws and regulations, traffic signs, contract, legal, etc.	Business letter, invoice, manual, contract, name card, packaging, etc.	Textbook, academic journal, dictionary, scientific communication, etc.

Based on EU's "Common European Framework of Reference for Languages" (2004)

Different language works in different domains

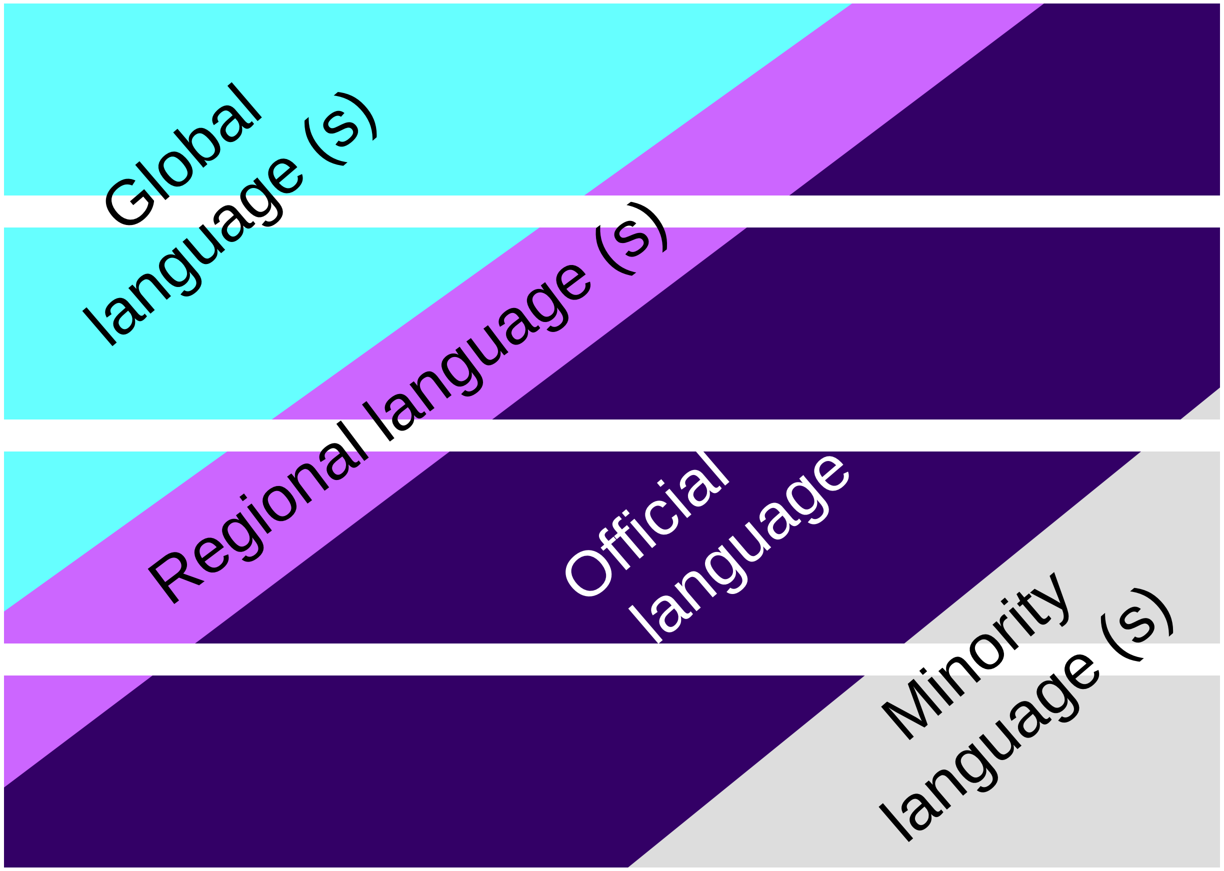
Socio-cultural domains

ac.xx
educational
domains

com.xx
occupational
domains

gov.xx
public
domains

others
personal
domains



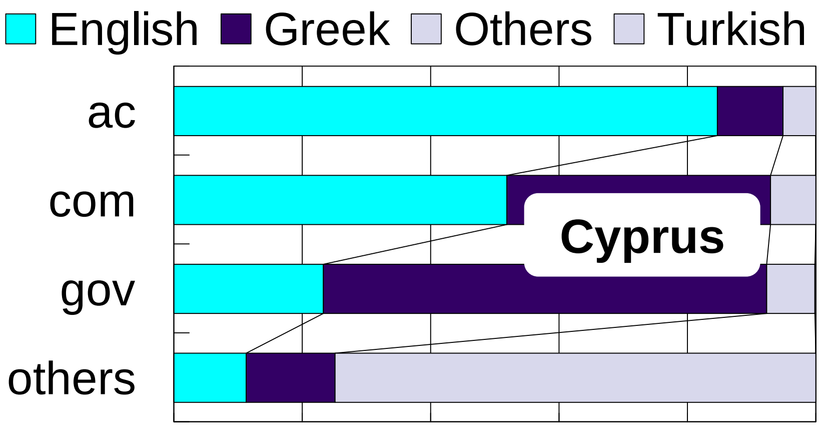
Global language (s)

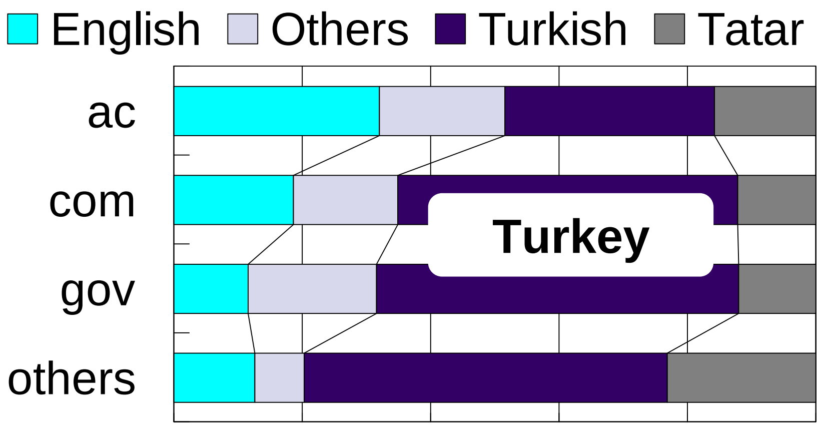
Regional language (s)

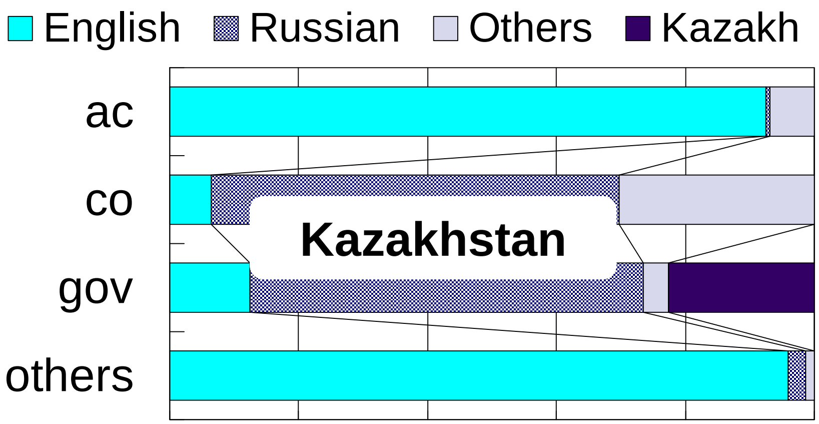
Official language

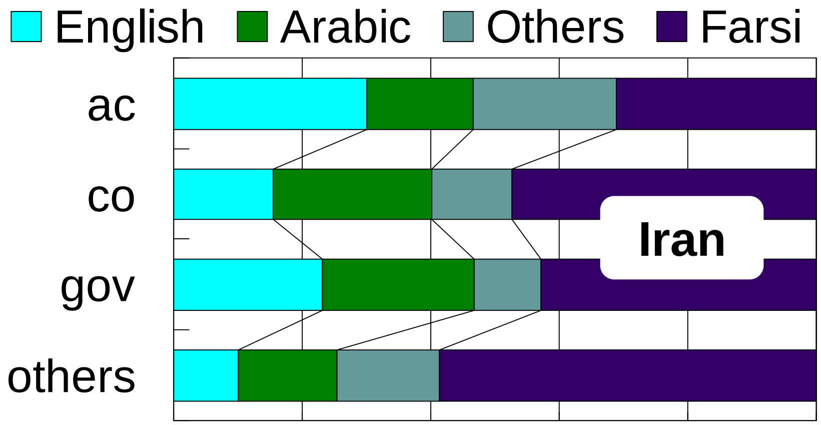
Minority language (s)

Specialization of Language Secondary domain analysis









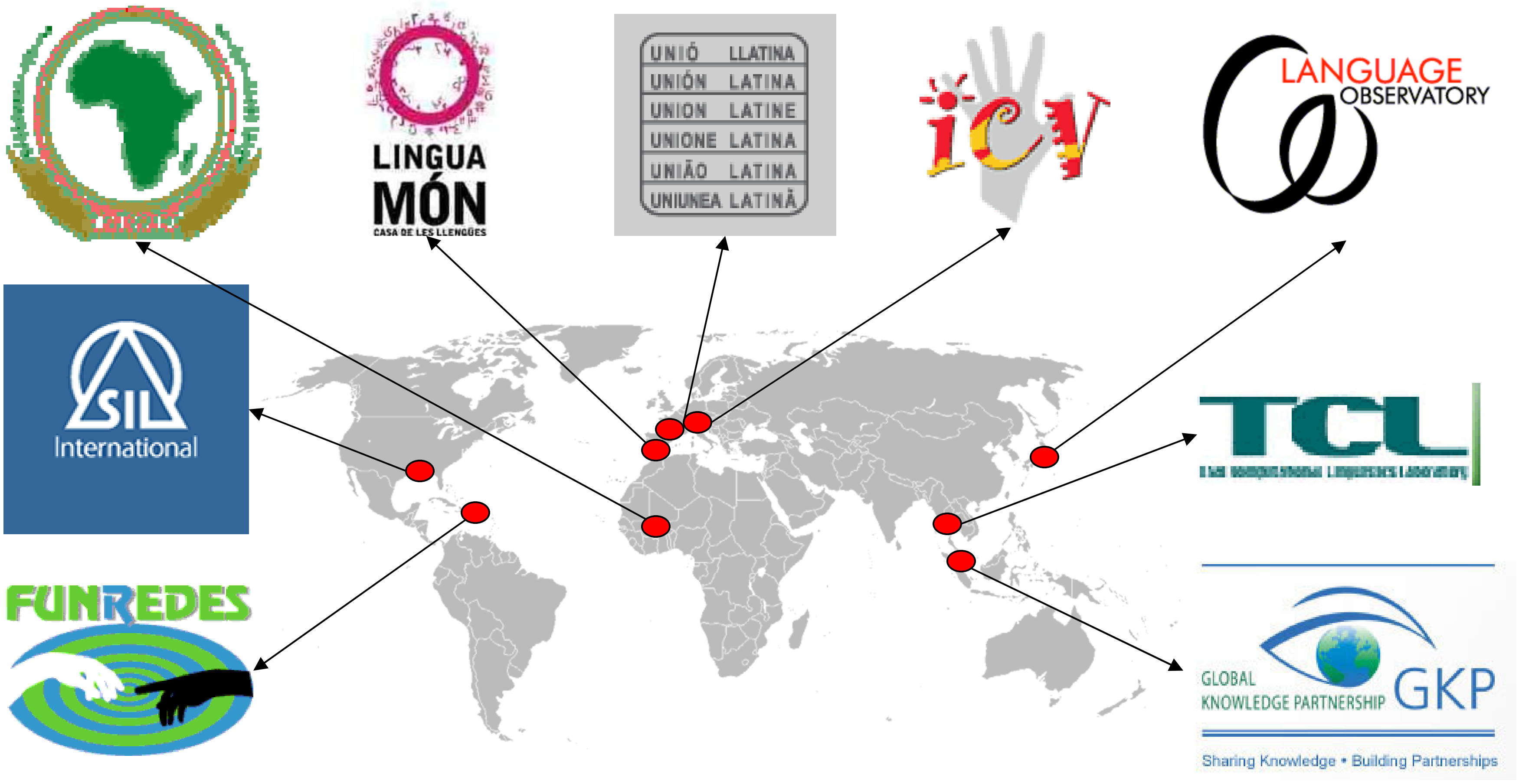
4. Conclusion

- “Digital Language Divide” observed
  - Economic context: Access opportunity divide
  - Technical context: Localization delay
  - Socio-cultural context: Empowerment of Mother Languages is needed

- Future of Language Observatory
  - Language-specific search engines
  - Language Observatory Network

World Network for Linguistic Diversity

maaya



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

Jehan Rectus Square, Paris
photo: courtesy by Wunna Ko Ko, June 2005