Joint UNESCO and ITU
Global Symposium on Promoting the Multilingual Internet

Free Open Source Software for facilitating language flows

Dawit Bekele

Assistant Professor, Addis Ababa University

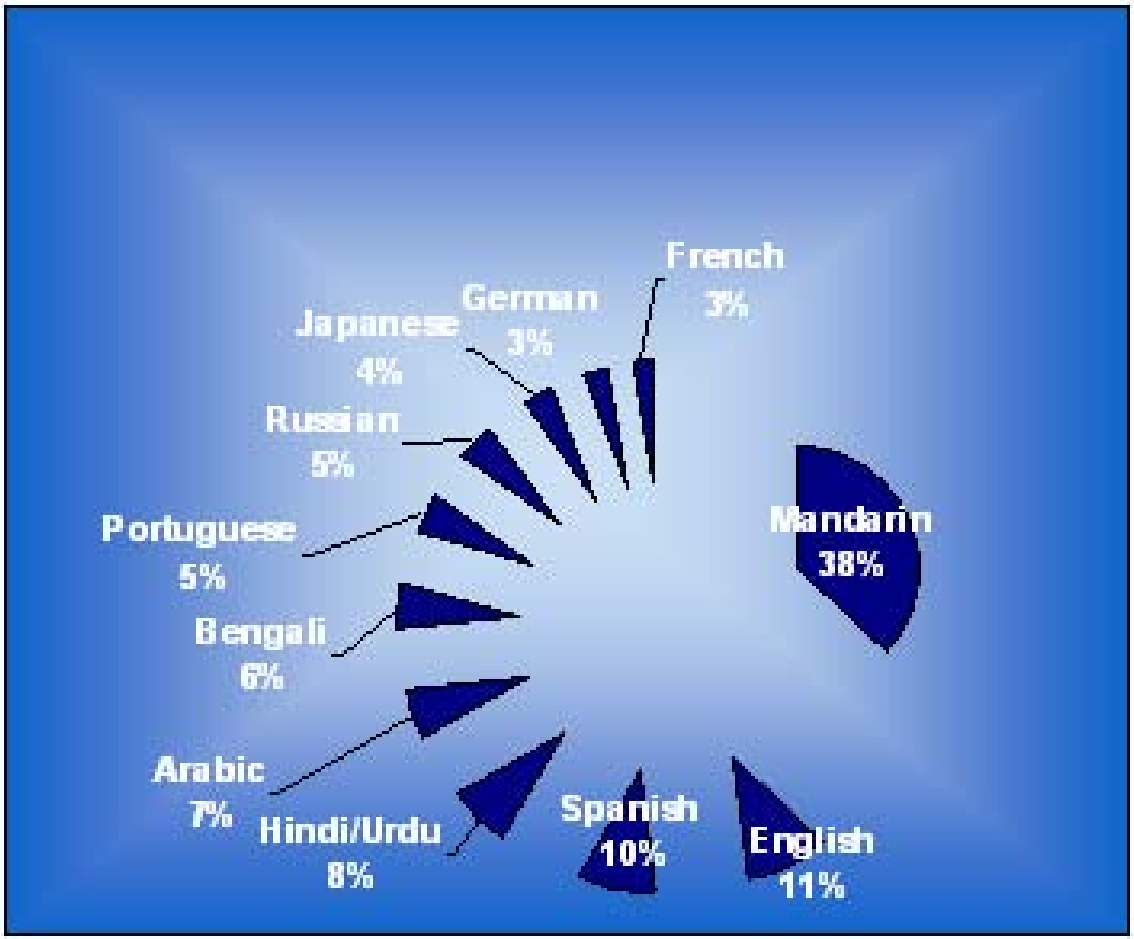
Coordination, Ethiopian Free and Open Source Network

Geneva, 9-11 May 2006

Introduction

- Multilingualism is increasingly considered as very important for the stability, peace and development in the world
- It should be nurtured in the real world as well as in the cyberspace
- However, the cyberspace tend to be dominated by some “major” languages
- This presentation tries to
  - show the discrepancy that exists in terms of the languages in the world and the cyberspace
  - Identify hardware and software causes
  - Propose solutions

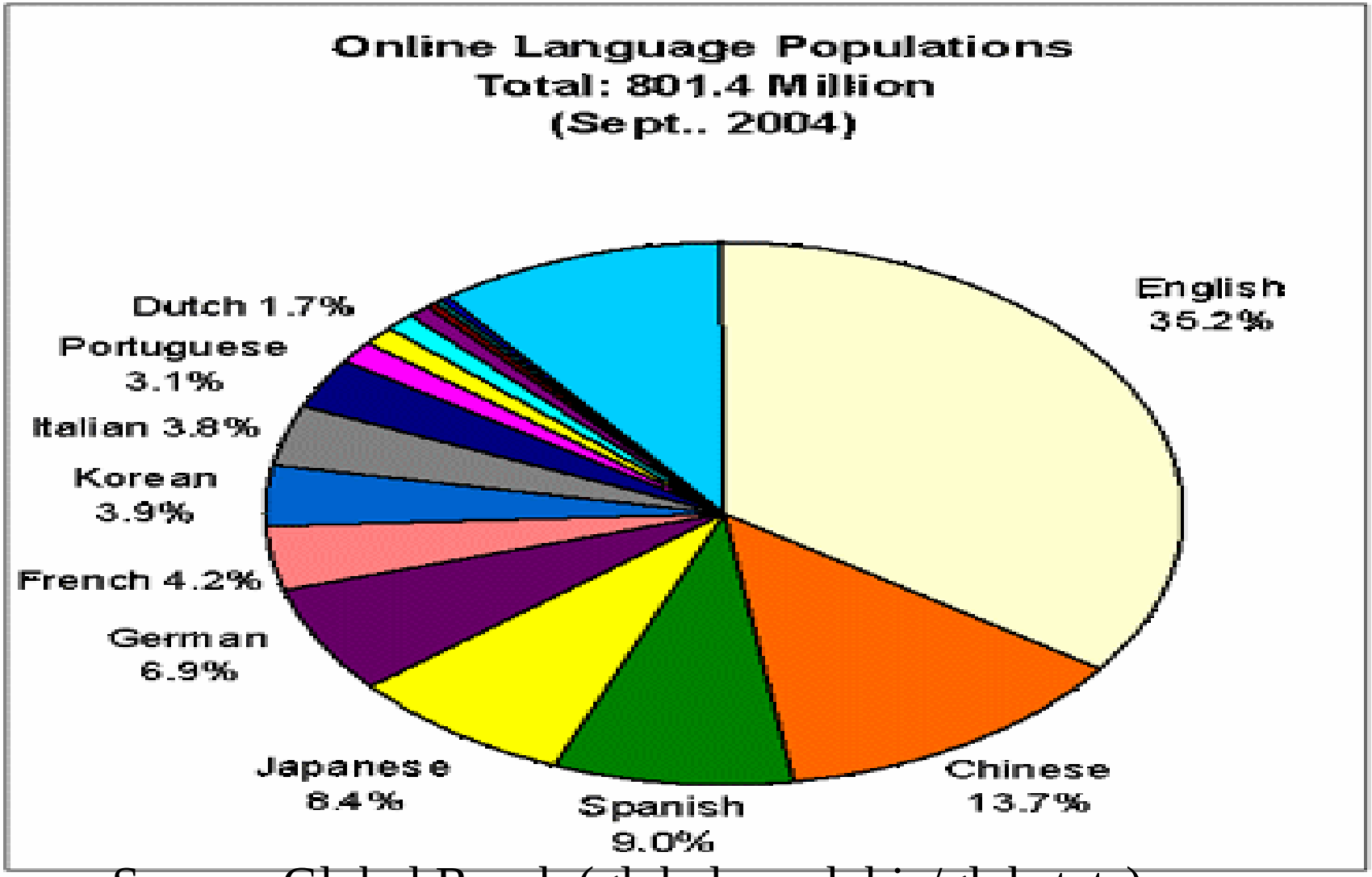
Languages of the world



Languages of the world

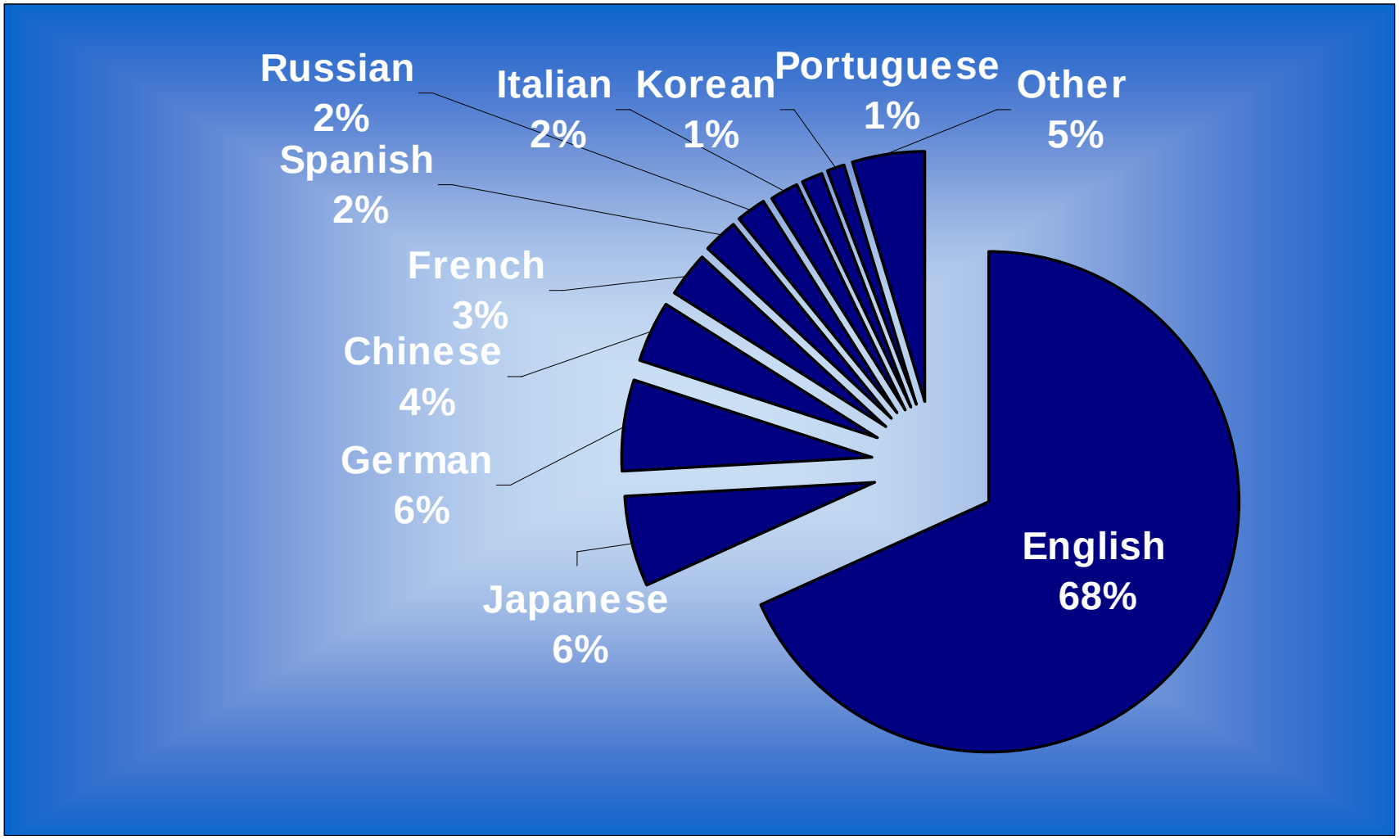
Language	Population
Mandarin	1,100,000,000
English	330,000,000
Spanish	300,000,000
Hindi/Urdu	250,000,000
Arabic	200,000,000
Bengali	185,000,000
Portuguese	160,000,000
Russian	160,000,000
Japanese	125,000,000
German	100,000,000
French	75,000,000

Languages on the cyberspace – On-line population



Source: Global Reach (global-reach.biz/globstats)

Languages on the cyberspace - Content



Source: Global Reach (global-reach.biz/globstats)

Problems

- Many languages do not exist in the cyberspace
  - **90%** of languages are not found on the Internet [Source: http://alumni.indiana.edu/lostlanguages/stats.html]
  - Thousands of languages worldwide are absent from Internet content [Source: UNESCO]

Problems …

- The proportion of language of the content on the cyberspace is very different from that of the population
  - English
    - 11% of the population of the world speaks English
    - 35% of on-line population has English as a native language
    - 68% of the cyberspace content is in English
  - Mandarin
    - 38% of the population of the world speaks Mandarin
    - 13.7% of on-line population has Mandarin as a native language
    - 4% of the cyberspace content is in Mandarin
  - There are many languages of the world where the situation is much worse than Mandarin

Why this discrepancy?

- Economic
  - Digital divide
- Technological
  - The focus of this presentation
- Two main technological problems
  - Problem to develop the digital content in the languages using computer hardware and software
  - Problem of representation of content

Problem to develop the digital content

- Difficulty to find hardware that supports some languages' scripts (ex. Keyboard)
- Difficulty to find the software to develop content in the language of the content
  - Example: Ethiopic content has been developed for decades with software with English interface
- Limits the number of people who can develop content in that language

Problem to develop the digital content …

- What is the solution?
  - Localization – change the interface of the software to local languages, culture and tradition
  - For decades, proprietary software developers didn't want to localize for languages that do not have economic power
    - Example: Microsoft just started to be interested in African languages
  - Recently, many localizations are being done using Free and Open Source Software

Problem to develop the digital content …

- What is the solution …
  - Free and open source software (FOSS) gives the freedom to
    - Copy
    - Distribute
    - …
  - FOSS provides the source code that anybody can modify
  - FOSS is gaining a lot of popularity around the world
  - A lot of software have been localized thanks to FOSS

FOSS gives the freedom to localize

- No need of authorization to localize a FOSS software
- Economic reasons not to localize are much less important than with proprietary software
  - The developers are not necessarily the localizers
  - Localizers have other reasons (pride, political will, technical interest, etc.)
- Since the source code is available all sorts of localization are possible
  - Time and date localization
  - Customization

Problem of representation of content

- There are hundred of alphabets and scripts used to represent the content in the languages of the world
  - Latin for Western European
  - Cyrilic for Eastern European
  - Ethiopic for Ethiopia and Eritrea
  - ...
- Until the 1990s ASCII, was the standard of Internet and was adequate only for Latin based languages
- It was necessary to use complex methods just to represent the content of other scripts and alphabets
  - Ex: Amharic: Image, Specific downloadable font, etc.
- UNICODE is solving the problem of encoding since it is the CODE of the world scripts and alphabets
- But there are other standards that need to consider multilingualism (ex. XML) in order to be able to develop a content in any language with the same ease as in English

Conclusion

- The first step towards multilingualism in the cyberspace is to have, in the cyberspace,
  - content in all languages of the world
  - in a proportion that respects the population of the world
- The use of FOSS can help achieve this objective since
  - it facilitates localization,
  - which in turn facilitates development of content in local languages