

Language enablement

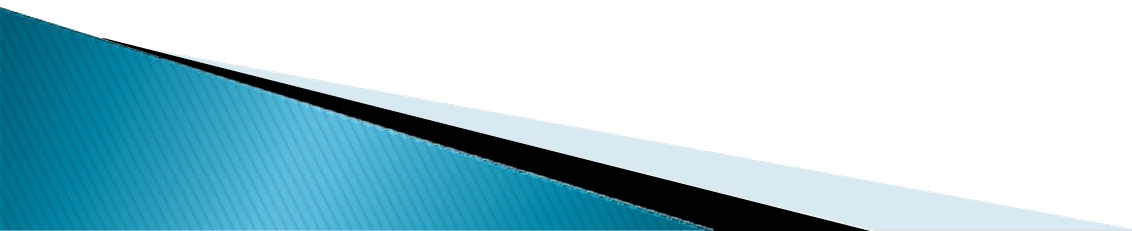
Andrew Cunningham

Don Osborn

Wunna Ko Ko

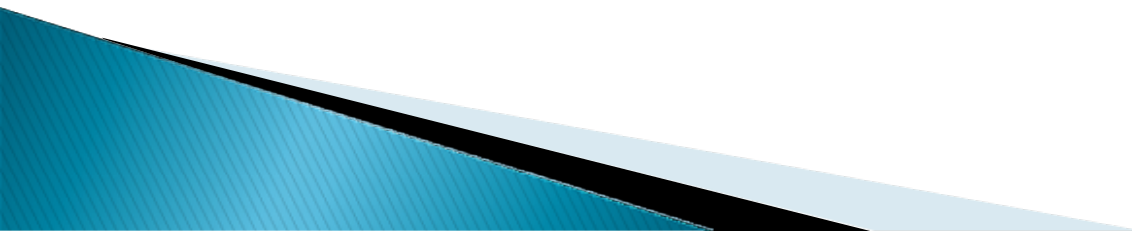


Language enablement

- Many languages have limited or no support within commercial operating systems or applications
 - The aim of language enablement is to ensure that a language is adequately supported
- 

Minimum requirements

- Display text correctly
 - Font rendering/text layout (including line breaking algorithms)
 - Fonts
- Type in language
 - Keyboard layouts and IMEs

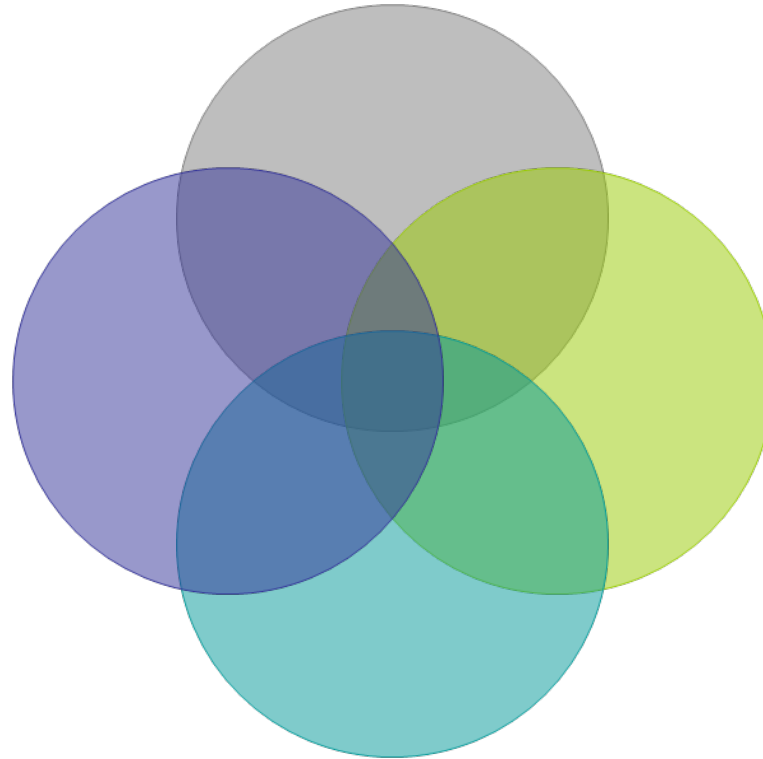




Text and font
services

Tools and
localization

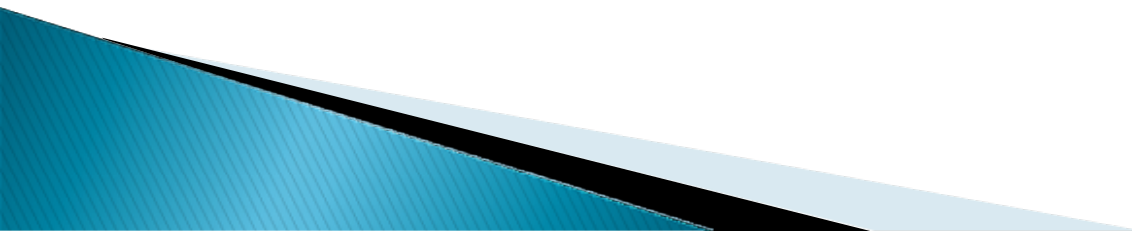
Locale
support



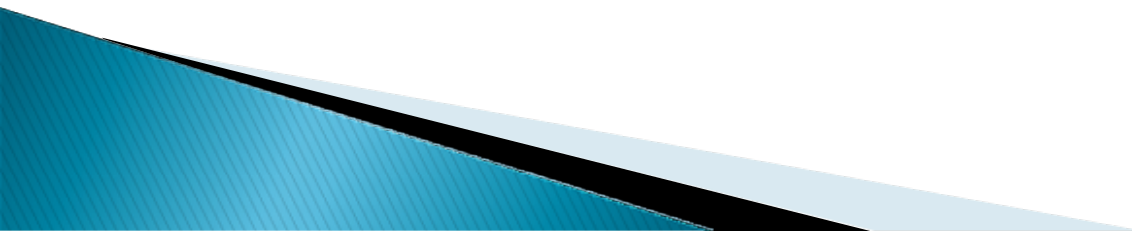
internationalization
architecture



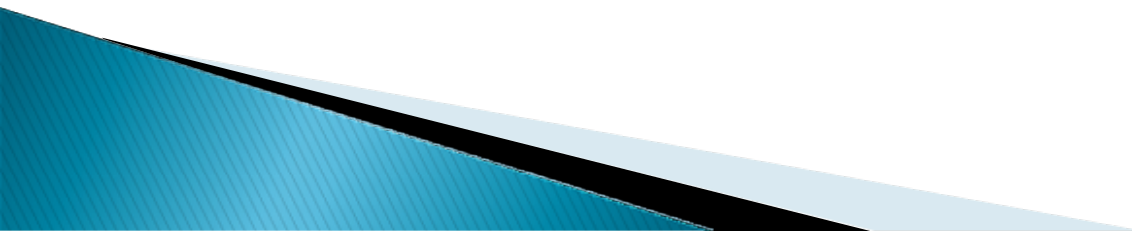
Australian context 1

- ▶ Changing migration patterns
 - ▶ New communities and new languages
 - ▶ Need for translations for print and web
 - ▶ Need for scripts for audio and video productions
 - ▶ Need for language ICT support for cultural and linguistic preservation, maintenance and development
- 

Australian context 2

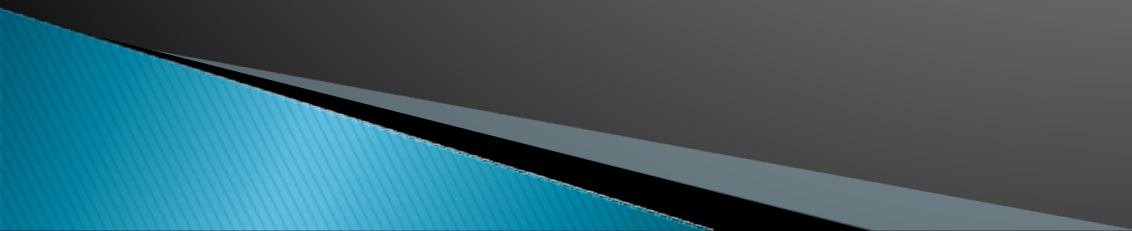
- ▶ New and emerging languages not supported by operating system and application vendors
 - ▶ May have high IT overhead
 - ▶ Assumption that solutions are already available, someone overseas has done the hard work
- 

Australian context 3

- ▶ Little or no language enablement work occurs within Australia
 - ▶ Some exceptions
 - Tavultesoft
 - Aboriginal language projects
 - Open Road
 - Individual linguists
- 



African languages

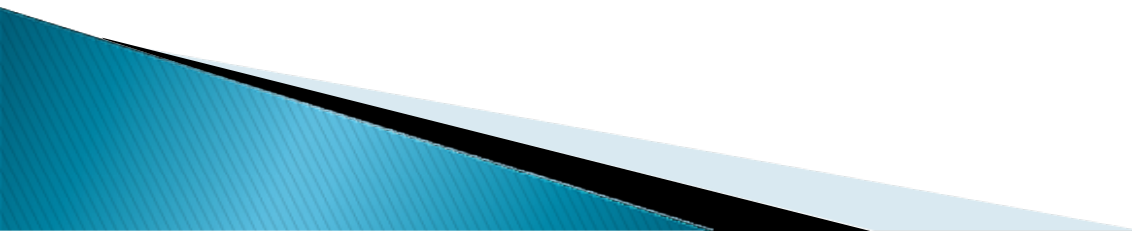




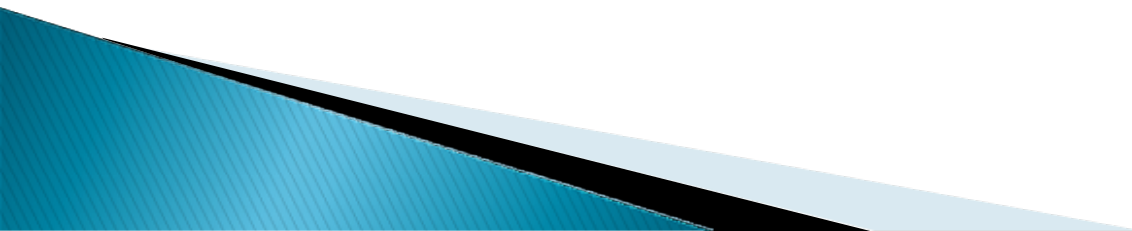
African overview

- ▶ Characterised by its cultural, linguistic, and geo-political diversity
- ▶ Linguistic legacy of colonial history
 - European languages as national languages
 - Language communities divided
- ▶ Multiplicities
 - Languages not confined to national borders
 - Orthographies may change between region or countries

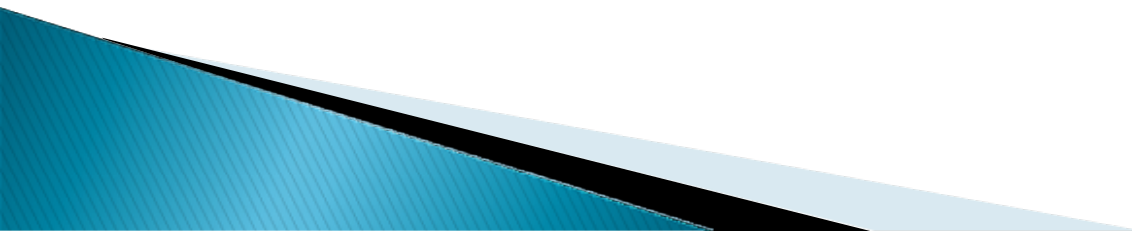
Writing scripts

- ▶ Latin (simple or complex)
 - ▶ Arabic
 - ▶ Ethiopic
 - ▶ Tifinagh
 - ▶ N'ko
 - ▶ Vai
- 

Barriers 1

- ▶ Low/unreliable electricity supply
 - ▶ Limited access to computers
 - ▶ Older computers
 - ▶ Low/poor internet connectivity
 - ▶ Literacy levels
 - ▶ Language policies give low priority to local languages
 - ▶ Most languages only have a short history of being written languages
- 

Barriers 2

- ▶ Reliance on former colonial languages for government and education
 - ▶ Sometimes the Latin script needs to be treated as a simple script.
 - ▶ Lack of knowledge and awareness of how Unicode works.
 - ▶ Lack of government standardisation and policies: keyboards, ICT policies, local language policies
- 

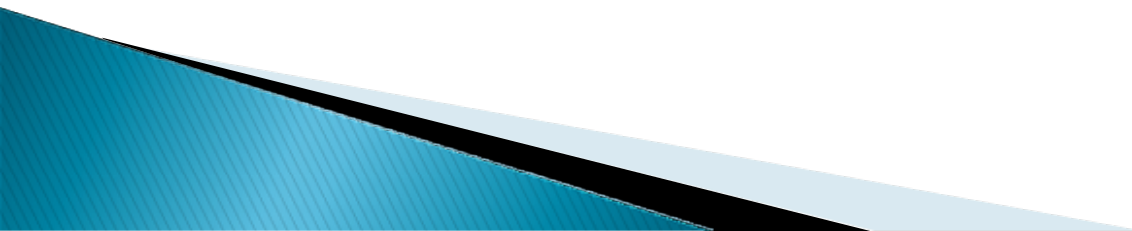


Actors

- ▶ Individuals
- ▶ Small open source localisation teams
- ▶ Not for profits
- ▶ Some small businesses
- ▶ African Localization Project – IDRC funded

Myanmar

Myanmar overview

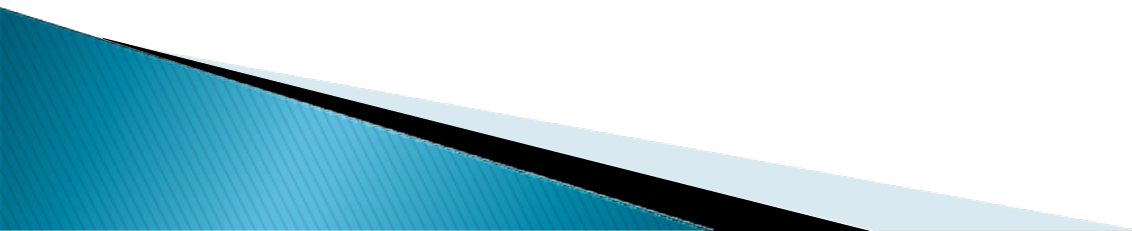
- ▶ Burmese is the official language of the Union of Myanmar, spoken by about 35 million of people as their mother tongue.
 - ▶ Burmese, Shan, Mon and the Karen languages are written in the Myanmar script.
- 



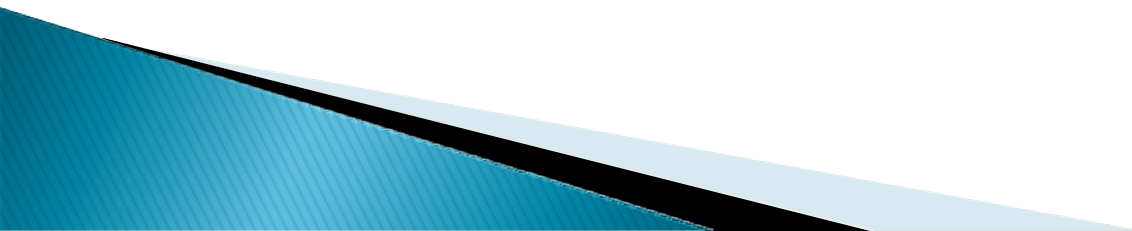
Status

- ▶ Supported by Unicode
- ▶ Not supported by major operating system and application vendors, i.e. Microsoft, Apple, Sun.
- ▶ Politics of the initial Unicode Myanmar block
- ▶ Limited ICT4Dev involvement
- ▶ Progress driven by individuals and small teams of developers.


Internal barriers

- ▶ Lack of standardization by Govt. and Govt. sponsored organisations, e.g. no national standards for character sets, keyboard layouts, etc.
 - ▶ Knowledge and skill gaps: some local developers and government agencies lack knowledge about benefits and potential for Unicode, how Unicode handles Myanmar, and the role Unicode plays in many modern operating systems and applications
- 

Myanmar script - evolution

- Legacy encodings
 - Still using in many areas of industry, especially in publication industry
 - Pseudo-Unicode solutions
 - Unicode 4 + UTN 11
 - Unicode 5.1 + UTN 11 (v2)
- 

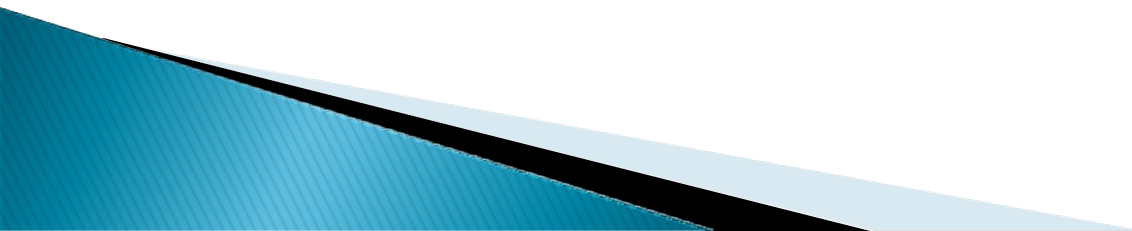
Rendering and text layout issues

- ▶ Complex rendering
 - Positioning of vowel markers, stacking consonants, etc.
 - Requires specialised rendering system: Uniscribe, Pango, Graphite, ICU Text Layout.
 - ▶ Line breaking: Myanmar text does not use white space to separate words. Spaces are used to distinguish between phrases and sentences.
 - ▶ Word segmentation
 - ▶ Collation/sorting
 - ▶ No Fonts suitable for UI
- 

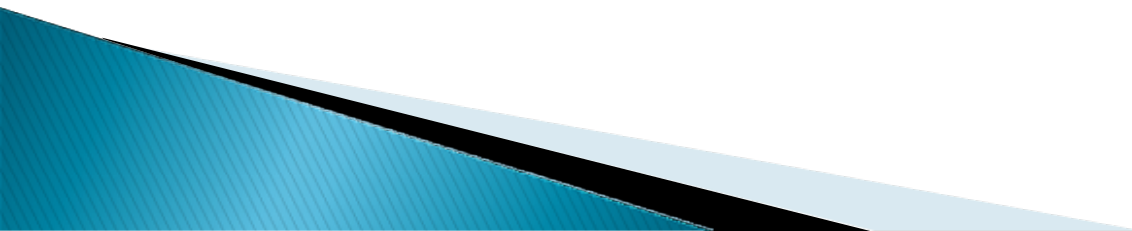
Pseudo-Unicode

- Used Unicode characters as a starting point
- No support for complex rendering
- Add additional glyphs as separate characters. Added characters to empty positions of the Myanmar block and to the Unicode PUA
- Possible conflicts with other uses of PUA or new Unicode character assignments
- Essential old legacy fonts masquerading as Unicode.
- Fonts from different vendors are not interoperable.
- Incompatible with new Unicode 5,1 based applications.
- Found a lot of usage on Blogs.

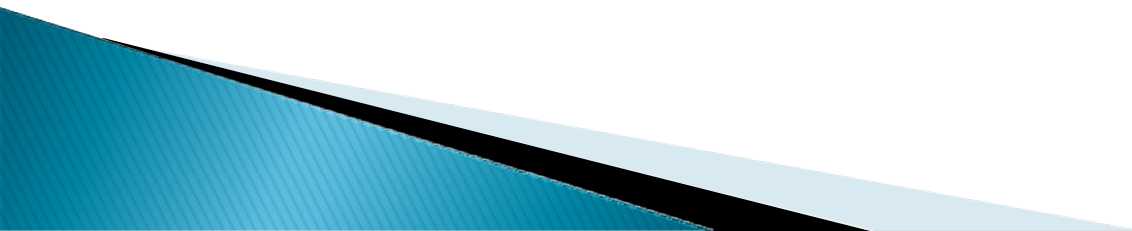
Unicode 4/5

- Only supports Burmese language
 - UTN 11 documented finer points of supporting Myanmar script
 - Unable to implement full Burmese language features
 - Not compatible with Unicode 5.1 solutions
- 

Unicode 5.1

- Fundamental changes to medial characters, Asat, and Virama for Burmese
 - Additional characters for Mon, Shan and Karen languages
 - New projects using Unicode 5.1
 - *The way forward for Myanmar computing.*
- 

Barriers for ethnic languages

- ▶ Although extensions for ethnic languages have already been added, there are few implementations under development.
 - ▶ No implementation guidelines (like UTN 11 for Burmese) have been published.
- 

Unicode 5.1 based Burmese projects

- ▶ Open source OpenType and Graphite fonts. (still very few)
 - ▶ Open source input solutions using SCIM+KMFL on Linux. (still very few)
 - ▶ Firefox and Thunderbird language packs
 - ▶ CLDR my-MM locale under development
 - ▶ ICU collation and line breaking algorithms
 - ▶ Burmese localization of OpenOffice
- 