Technical and Standardisation Challenges for Nepalese Writing

Pat Hall

Language Technology Kendra

Nepal

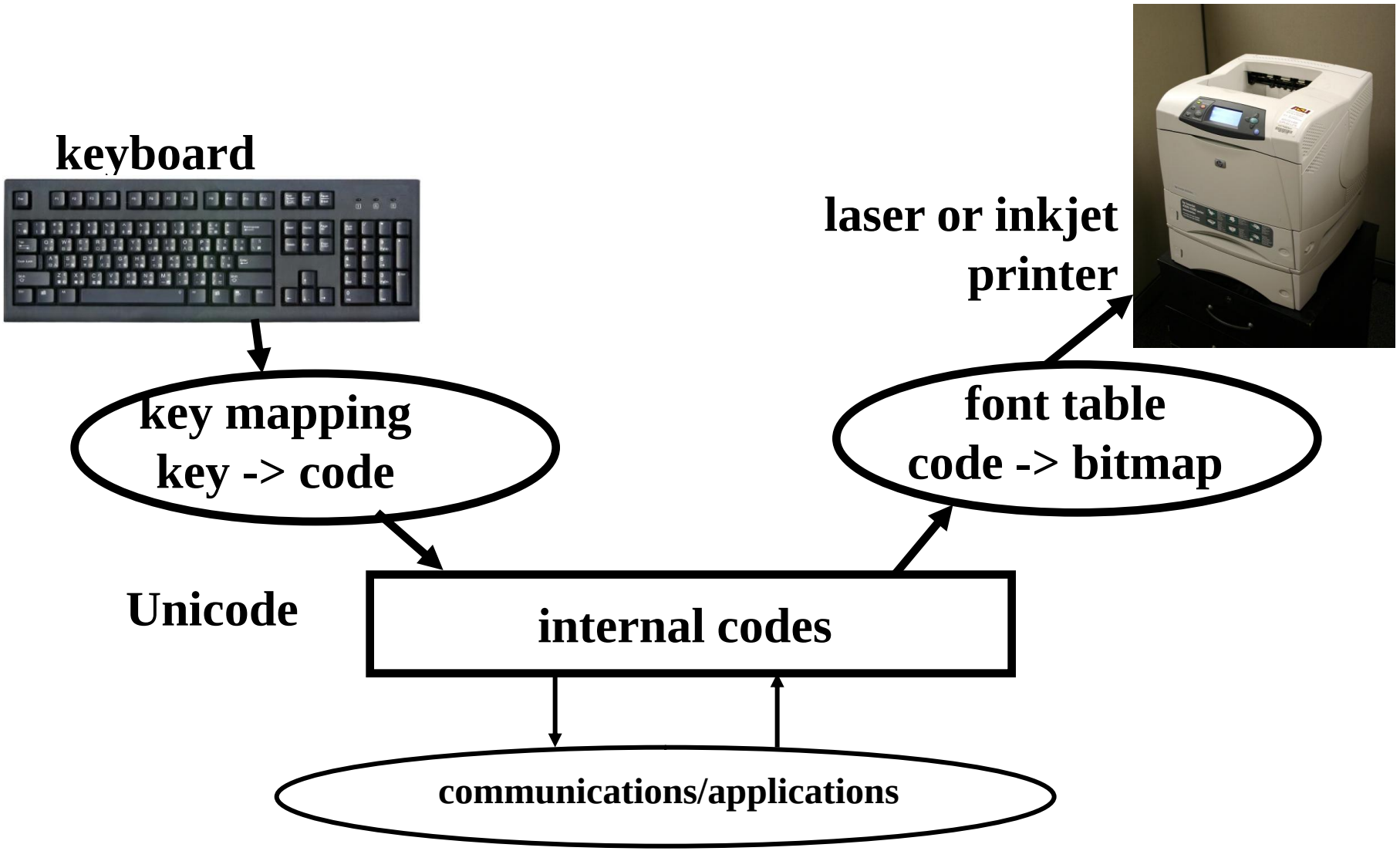
the changing way we write

the changing way we write

the changing way we write

the changing way we write

but what matters is



but what matters is

laser or inkjet printer

font table
code -> bitmap

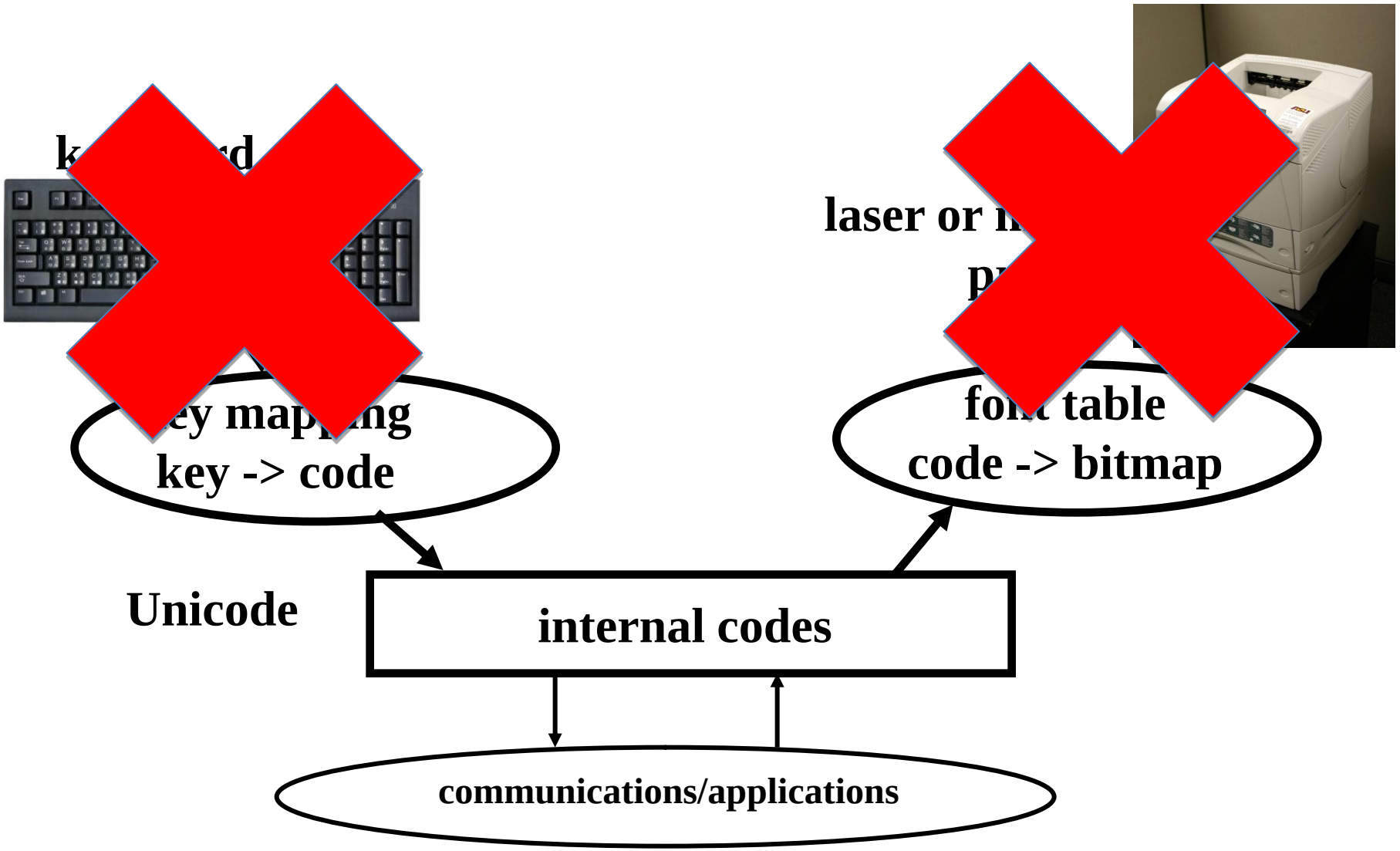
key mapping
key -> code

Unicode

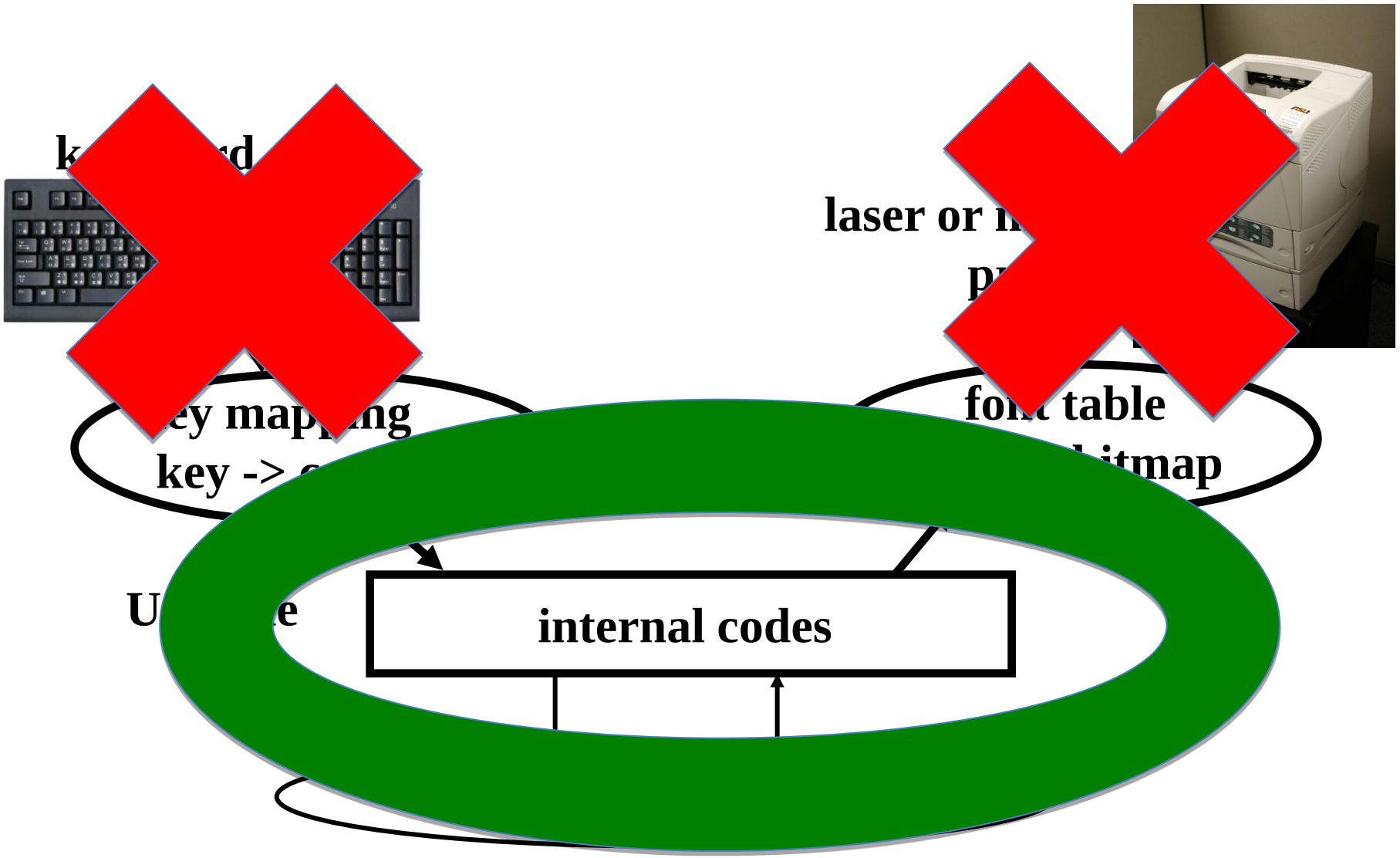
internal codes

communications/applications

but what matters is



but what matters is



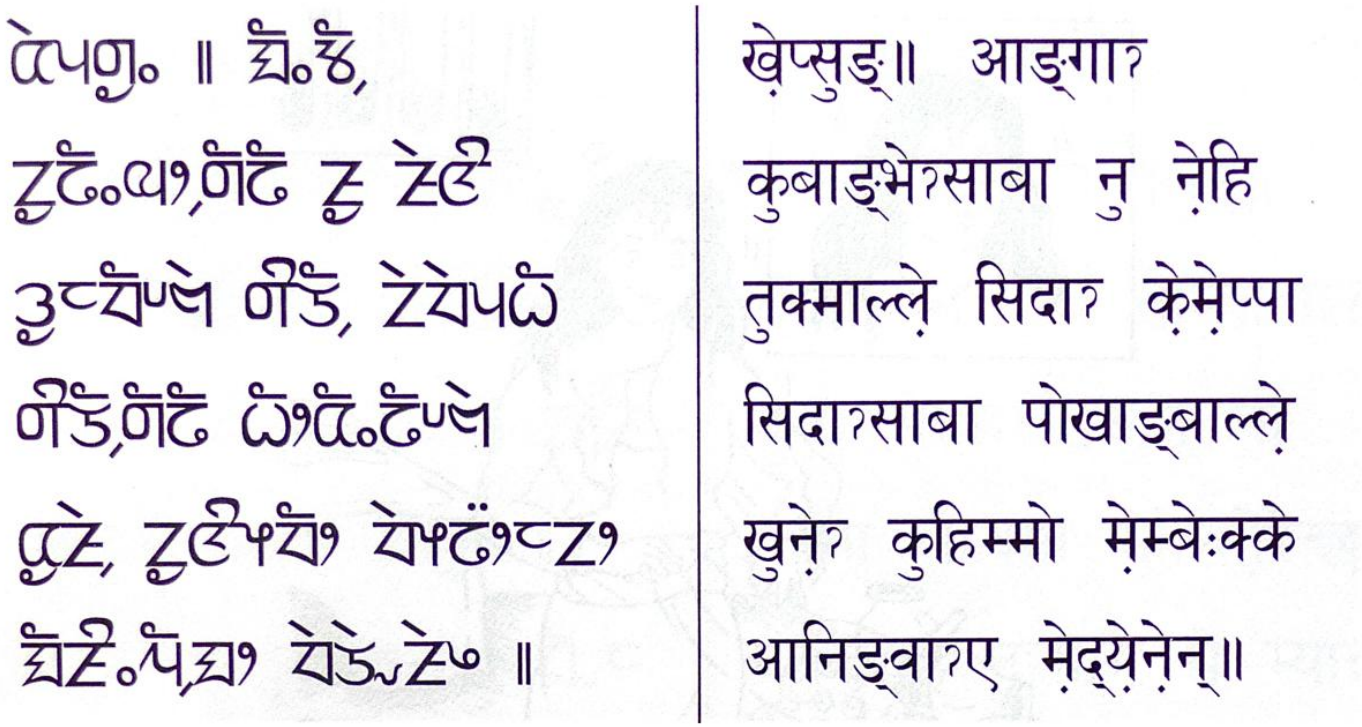
deep structure characters

Example 1. Limbu

- written in Devanagari under one language policy
  - people know this
  - technology available
  - glottal stop added to Unicode Devanagari
- but had own writing system Sirijanga
  - created 18th Century, revived 1920s
  - Unicode table 2002 by Michailovsky and Everson

	190	191	192	193	194
0	ᤀ	ᤐ	◌ᤰ	◌ᤵ	᥆
1	ᤁ	ᤑ	◌ᤱ	◌ᤶ	
2	ᤂ	ᤒ	◌ᤲ	◌ᤷ	
3	ᤃ	ᤓ	◌ᤳ	◌ᤸ	
4	ᤄ	ᤔ	◌ᤴ	◌᥆	᥇
5	ᤅ	ᤕ	◌ᤵ	◌᥇	᥈
6	ᤆ	ᤖ	◌ᤶ	◌᥈	᥉
7	ᤇ	ᤗ	◌ᤷ	◌᥉	᥊
8	ᤈ	ᤘ	◌ᤸ	◌ᥚ	᥋
9	ᤉ	ᤙ	◌᤹	◌᥋	᥌
A	ᤊ	ᤚ	◌᤺	◌᥌	᥍
B	ᤋ	ᤛ	◌᤻	◌᥍	᥎
C	ᤌ	ᤜ			᥏
D	ᤍ				ᥐ
E	ᤎ				ᥑ
F	ᤏ				ᥒ

Limbu text in both writing systems



one-to-one correspondence

one language two writing systems - could we have one encoding of the language with two fonts?

one language multiple writing systems

not unprecedented but unusual

- Turkish
  - Arabic to Roman
- Central Asia
  - Arabic to Cyrillic to Roman
- Sindhi and Punjabi
  - in Arabic in Pakistan
  - in Devanagari in India

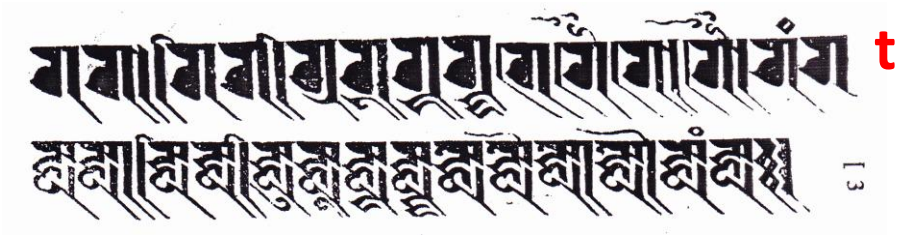
encode the language
the phonemes
not the writing!
re-map fonts

Example 2. Newari / Nepal Bhasha

- oldest written tradition in Nepal.
- 1990s hack fonts produced
- 2000 Unicode proposal drafted by Everson
- 2008/10 Unicode drafts from Newars
- meetings of Nepal Lipi Guthi
- 2010 Ranjana Unicode project driven by northern antiquarian interests.

2010 Unicode Ranjana project

“Since Ranjana is visually and structurally similar to the Lantsa and Wartu scripts used for Buddhist Sanskrit documents in Tibet (China), Bhutan, Mongolia, Nepal, Sikkim and Ladakh (India) it has been considered would be practical to merge these two scripts (Lantsa and Wartu) with Ranjana for encoding purposes.”



the emphasis was on the writing, with much reference to antiquarian texts

the Newars pleaded for Prachalit, and after much fierce debate, the project ended within a year with no proposal

“Analysis of the chief features suggests that a number of these can be unified on structural grounds.”

Analysis of Rabison Shakya's book

seven more!

prachalit

why?

ranjana

bhujimmola

conjuncts?

conjuncts?

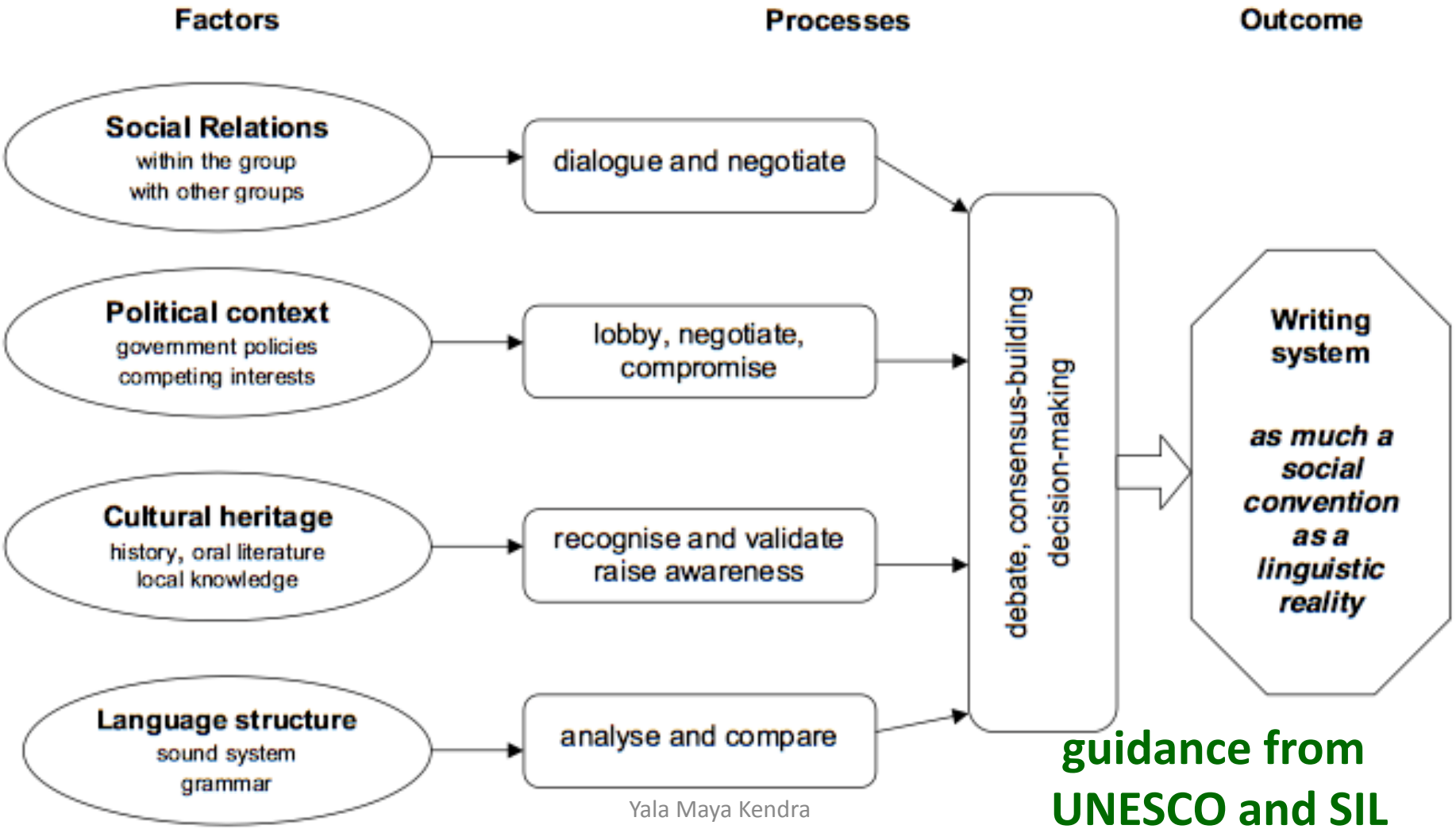
prachalit rearranged in vargs

Aspirated nasals! Not in Ranjana or Bhujimmola! Glyphs look like conjuncts! ***Are they conjuncts?***

some possible conclusions

- one language with many styles of writing
  - cannot map onto Devanagari
    - because more phonemes
  - but needs checking using contrastive pairs
- need a distinctive single unicode block
  - and many fonts, at least one per style

Example 3. Unwritten languages



guidance from UNESCO and SIL

central stages

1. determine phonemic inventory
2. map to some writing system(s)
   - new glyphs as necessary
   - usually locally dominant system
   - develop spelling system
   - encode writing
3. OR encode phonemes directly
   - direct link to speech
     - TTS reads the text to people
     - ASR "writes" the speech
   - can write in many scripts
   - standarize codes and mapping to characters

In Nepal mostly based on Devanagari
Michael Noonan survey
Bhim Regmi's proposal

Lohrung request for Roman

but many want to write in Roman why not write in both?

नेपाली भाषा नेपाल बाहेक भारतमा पनी

nepālī bhāṣā nepāl bāhek bhārtmā pni

संवैधानिक हैसियत प्राप्त भाषा हो । नेपाल,

s˜vaidhānik haisiyt prāpt bhāṣā ho. nepāl,

भारत र संसारका अरू वभिन्नि स्थानमा छरिएर

bhārt r s˜sārkā arū vibhinn sthānmā chrier

diacritic needs positioning

implicit short vowel missing

the Romanagari transliteration is rendered directly from the Devanagari codes following the Library of Congress ALA-LC romanization standards

Example 4. Gurung and Magar

designs for new writing

Yala Maya Kendra

what is going on?

Mark Turin

marking identity by being visually distinct

could we mark language identity in font variants?

What should be standardized?

- separate code sets for each language
  - + separate identity for each language
  - - will require around 5,000 code points
  - – extra cost per language
- one, or a few, unified code sets for Nepal
  - + fewer standards to write
  - + can share cost of technology
  - – agreement required over many communities
- mapping from codes to characters

how should they be standardised?

- Nepal Standards Organisation
  - ✓needs to be broadened to mirror ISO
  - ✓must reach agreement within Nepal first
- International Standards Organisation
  - ✓need champion member
    - India?
  - ✓and champion expert
    - somebody who knows Nepal
  - ✓attending meetings important but costly

Thank You

key issues for writing languages

- **the phonemes and their computer codes**
  - **some way of writing these for debate**
  - **standardize mapping phonemes to characters**
- **how to reach agreement**
  - **enshrine that in standards**