Negotiating the issues of encoding and producing traditional scripts on computers: Working with Unicode

Deborah Anderson
Script Encoding Initiative
Department of Linguistics, UC Berkeley

and

Stephen Morey
Centre for Research on Language Diversity
La Trobe University
How Unicode Process Works

1. Users, linguists identify script/characters not in Unicode/ISO standard
2. Unicode script proposal written
3. Two standards committees review proposals and vote whether to accept them
4. Publication of script in Unicode/ISO standard
5. Create fonts, keyboards, update software
Script diversity in South and South East Asia

- Indic-based (Brahmi-derived)
  - Thought to originate from contact with West Asia/Europe
- ‘Indigenous scripts’
  - Chinese
  - Japanese Kana
  - Korean
- New Scripts
Writing systems of the world
Case Studies

1. Tai in Northeast India
   - Grouping glyphs regarded by speakers as different in the same encoding points

2. Assamese / Bengali
   - Naming of characters

3. Ordering for Tai words
   - Visual order versus Logical order
1. Tai in Northeast India (a)

Two scripts in Northeast India

(1) **Tai Ahom** script, said to have been brought from Muang Mao in the 13th century. Historically very similar to Dehong Dai or Tai Mao script

(2) **To Lik Tai** (body book Tai) – scripts for the spoken Tai languages based on Shan and Burmese, used for Aiton, Phake, Khamti, Khamyang
1. Tai in Northeast India (b) Tai Ahom
1. Tai in Northeast India (c)
To Lik Tai and Burmese
1. Tai in Northeast India (e) Comparison

<table>
<thead>
<tr>
<th>Letter</th>
<th>Tai Mau</th>
<th>Tai Ahom</th>
<th>To Lik Tai</th>
<th>Padauk</th>
<th>Burmes e</th>
</tr>
</thead>
<tbody>
<tr>
<td>ka</td>
<td>□</td>
<td>ꈨ</td>
<td>ꈨ</td>
<td>ꈨ</td>
<td>ꈨ</td>
</tr>
<tr>
<td>kha</td>
<td>□</td>
<td>ꈨ</td>
<td>ꈨ</td>
<td>ꈨ</td>
<td>ꈨ</td>
</tr>
<tr>
<td>nga</td>
<td>□</td>
<td>ꈨ</td>
<td>ꈨ</td>
<td>ꈨ</td>
<td>ꈨ</td>
</tr>
<tr>
<td>cha</td>
<td>□</td>
<td>ꈨ</td>
<td>ꈨ</td>
<td>ꈨ</td>
<td>ꈨ</td>
</tr>
<tr>
<td>sa</td>
<td>□</td>
<td>ꈨ</td>
<td>ꈨ</td>
<td>ꈨ</td>
<td>ꈨ</td>
</tr>
</tbody>
</table>
1. Negotiating with Unicode

1. Separately encoding a script
   - Structural differences or just different glyph shapes?

2. Possibilities for To Lik Tai
   - Add new characters?
   - Create To Lik Tai font (at current code points)?
   - Use Variation Selector to get wanted shapes?
   - Ask Facebook to support 2 fonts?
   - Add wording about To Lik Tai to Unicode Standard?
2. Assamese (a)

- Today Assamese and Bengali languages are written with the same script.
- The two scripts have a common ancestor, developing separately over centuries and re-converging.
- The Unicode encoding of this common script is called simply ‘Bengali’.
- Members of the Assamese community are upset that the whole script is named ‘Bengali’ without reference to them.
2. Assamese (b) History

5th century: Umachal rock inscription
13th century: proto-Assamese shapes
Middle ages: Three varieties Kaitheli (used by non-Brahmins), Bamuniya (used by Brahmins, for Sanskrit) and Garhgaya (used by state officials of the Ahom kingdom)
19th century: first Assamese script for printing
   Bengali and Assamese lithography converged to the present standard that is used today.
20th century: Unicode names all characters as ‘Bengali’
2. Assamese (c)

The following two letters are not used at all in Bengali, but have been given complicated names by Unicode:

- ব Bengali Letter ra with middle diagonal
  - Real name: Assamese letter ra
- ব Bengali Letter ra with lower diagonal
  - Real name: Assamese letter wa
2. Assamese (c)

क्ष pronounced [kʰjɔ] or sometimes just [kʰ], this letter is historically a conjunct consonant of [k] and [ʂ] (/kʂ/) but is a full consonant in Assamese, such as

ক্ষণ ‘measure of time equal to 4 minutes, a while’
2. Negotiating with Unicode

- Name of script/characters: Problem
- For Bengali/Assamese issue (Unicode 1.1, 1993)
  - Changes to Unicode prose section, webpages
  - Submit information to Common Locale Data Repository?
3. Ordering for Tai words

Consider the Tai word /kɛ/ ‘old’
In standard Thai it has to be encoded as

ฤ + ก กɛ
ɛ + k kɛ

But in Shan, Tai languages of Northeast India, it has to be encoded as

ꯗ + ṝ ṝɛ
k + ɛ kɛ
3. Negotiating with Unicode

- Two encoding models:
  - Thai follows a **visual model**: type the letters as you see them in left-to-right order (for Thai, for ex.)
  - Shan follows the **logical order**: dependent vowels follow consonant, even though they display before the consonant. Logical order is the default for Unicode.
3. Negotiating with Unicode

- Scripts used for Tai languages:
  - Thai (1993, Unicode 1.1) [visual order]
  - Lao (1993, 1.1) [visual order]
  - Tai Dam = Tai Viet (2009, 5.2) [visual order]
  - New Tai Lue (2005, 4.1) [changing to visual]
  - Shan = Myanmar (1999 +) [logical order]
  - Ahom (2015, 8.0) [logical order]
  - Tai Tham (2009, 5.2) [logical order]
  - Tai Le (2003, 4.0) [logical order]
3. Negotiating with Unicode

- Ex. of encoding model change: New Tai Lue (published 2005, Unicode 4.1)
  - Originally encoded in logical order
  - Changed to visual order in 2015, because main user community in China had data stored in visual order and fonts relied on storing data in visual order
4. Improving Relations with Unicode

- Make contact with at least one member of the Unicode Technical Committee (or active contributor to Unicode) early on
- Try to call in to UTC meetings on topics of interest (or attend meetings in person)
- Meet with UTC member or participant, if possible
Acknowledgements

Thanks to Martin Hosken for all his efforts on behalf of scripts in Asia

Deborah Anderson’s research is funded by a grant from National Endowment for the Humanities (USA), PR-50205015, and from the Google Research Award

Stephen Morey’s research is funded as part of a Future Fellowships (2011-14) awarded by the Australian Research Council