**ANDERSON — Negotiating**

<1 image>

**Negotiating the Issues of Encoding and Producing Traditional Scripts on Computers—Working with Unicode**

Anderson, D. and Morey, S.

Over the past 30 years, developments in computing mean that almost every script and writing system ever created can be coded on a computer and used on Facebook, mobile phones, and in emails, and large numbers of documents can be encoded, searched, and archived in a range of different scripts.

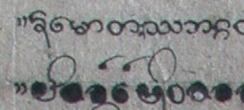
In South and Southeast Asia, there are a large number of different scripts, some used by quite small communities. In India, for example, there are around 11 official scripts, with states like Andhra Pradesh, Kerala, Gujarat, Karnataka, and Tamil Nadu having their own unique scripts. Speakers of Tai languages, in particular, have long had a range of different scripts in use among languages that are quite similar, so, for example, the script used by Shan people in eastern Myanmar is quite distinct from that of the Khün in the eastern part of Shan state, and also different from but intelligible with the Khamti script in western Myanmar, and the various scripts used by Tai people in Northeast India.

Since the earlier part of this century, a great effort has been made to encode all of these scripts in Unicode, a standard that allows for the encoding of symbols used in writing that can be demonstrated to be in use, or to have been in use in the past. However, negotiating a script into Unicode is a complex issue, involving considerable technical expertise and knowledge of script encoding principles, things that are difficult enough for an academic linguist but virtually impenetrable for members of the speech communities.

Combining our expertise in both script encoding and in linguistics, we will raise issues of community involvement in the process by means of several case studies.

1. The decisions to encode different forms of letters used in different scripts, such as Burmese on one hand and Tai varieties in Northeast India on the other, on the same encoding point so that they cannot both appear in platforms like Facebook (Hosken, 2014b).

Below is an example, from an 1821 Tai Phake manuscript called Mahosatha, telling one of the former lives of the Buddha (Jataka). This shows the stark differences in glyph shapes of the characters. The top line is in the Burmese script (Pali language), then shifts to Tai in the second line.



Background: In Unicode, the glyphs for the letters in the Burmese (‘Myanmar’ in Unicode) script were based on their use in the Burmese language, whose 32 million speakers dwarfs the Tai language communities of India (such as Phake with 2,000 speakers). The Myanmar script was published in Unicode in 1998, and the minority users were not consulted in the encoding then, so the default glyphs represent those of the Burmese language, not the Tai languages.

Update: The request (Hosken, 2014b) to Unicode to disunify the characters for Tai (Aiton and Phake) from Burmese was discussed in August 2014, with experts calling in to the Unicode Technical Committee meeting. However, no consensus was reached, so the topic was dropped for the time being. As a result, users on Facebook can’t show a mixture of Burmese and Tai-based languages on the same FB page with the expected glyph shapes.

2. The difficulties that the Assamese community find with the naming of letters used only in Assamese, such as ‘Bengali letter Ra with middle diagonal U+09F0’, a letter not used at all in Bengali but found in Assamese with the pronunciation [wɔ] (Hosken, 2012).

Background: The name of the script ‘Bengali’ was already the name of the script in Unicode 1.1 (1993). Due to stability policies, script names cannot be changed.

Update: The controversy reached the point of being discussed in the newspapers in 2012. In response to users, the Unicode Consortium made a few changes to incorporate ‘Assamese’ in the chapter on Bengali and the names list in 2012, but clearly there remains some confusion for Assamese users.

3. Problems for users posed by the Unicode New Tai Lue encoding model, which does not allow users to type the characters in visual order (Hosken, 2014a), the model adopted by the scripts used in neighboring countries of Thailand and Laos.

Update: As the result of the request to change the encoding model by Hosken (2014a) and feedback from a public review issue posted on this issue (Unicode Consortium, 2014a), Unicode decided to change its encoding model in the next version of Unicode, Unicode 8.0, which will be released in the summer of 2015.

The talk will conclude with suggestions on ways to improve the interaction between linguists, speech communities, and the Unicode standards committee. For example, conveying technical details over email and the phone can become problematic both for the user community and the Unicode Technical Committee (such as how to represent Tai languages on Facebook, above, 1). One strategy that has been successful is to bring a Unicode expert to the users (or to a linguist who works with the users) or to bring an expert to the Unicode committee. However, this involves hurdles, such as finding funding for travel.

The urgency for supporting the requirements of small language communities is based on the increasingly widespread use of mobile devices and social media. Why should minority language users in South and Southeast Asia be limited in their electronic communication, so that they have to rely on script-style of the larger language community, instead of having their own preferred style displayed?

**Funding**  
This work was supported by the National Endowment for the Humanities [grant PR-50205 to Deborah Anderson] and the Australian Research Council [Future Fellowship grant FT 1001006614 to Stephen Morey].

**References**

**Hosken, M.** (2012). Proposal for Minor, Non-Character, Additions to UCS Addressing Concerns from Assamese. Working Group Document, UTC L2/12-350,

http://www.unicode.org/L2/L2012/12350-assamese.pdf.

**Hosken, M.** (2014a). Proposal to Change the Encoding Model of New Tai Lue. Working Group Document, UTC L2/14-090, http://www.unicode.org/L2/L2014/14090-new-tai-lue.pdf.

**Hosken, M.** (2014b). Proposal to Disunify Khamti Letters from Myanmar (Revision 2). Working Group Document, UTC L2/14-108R2, http://www.unicode.org/L2/L2014/14108r2-khamti\_disunify.pdf.

**Unicode Consortium**. (2014a). Proposed Encoding Model Change for New Tai Lue. Public Review Issue #281, http://www.unicode.org/review/pri281/pri281-background.html.

**Unicode Consortium**. (2014b). *The Unicode Standard 7.0.0*. Mountain View, CA: Unicode Consortium, http://www.unicode.org/versions/Unicode7.0.0/.