L2/07-008

Workshop on Encoding and Digitizing Thai Scripts

Sponsored by UNESCO and the Vietnam Institute of Science
and Technology, Institute of Information Technology

Điện Biên, Vietnam

November 3, 2006

Report prepared for the

Script Encoding Initiative

by Jim Brase, SIL International
(revised Jan. 3, 2007)

Preconference meeting

Prof. Cam Trong of Hanoi University is one of the leading scholars of his people's writing system and language. Although he did not attend the workshop, a small group of us met with him in Hanoi on the evening of November 1.

Workshop and Participants

The workshop was held in Dien Bien Phu on November 3. It was attended by about 50 delegates. Kenichi Handa came from Japan, and James Do Ba Phuoc and I came from the United States. The other delegates were all from Vietnam, and included representatives from various agencies of government and technology, as well as representatives of the Tai language community.

Ngo Trung Viet and I visited Son La on November 4, and returned to Hanoi on November 5 by car.

Papers Presented

Several papers were presented in the morning session. The afternoon session included speeches by the delegates and discussion.

An asterisk (*) indicates papers included in the printed proceedings.

*The Viet Thai Scripts in the United States — Jim Brase, SIL International

*Viet Thai handling on GNU/Linux Systems — Kenichi Handa, AIST Japan

Chữ Thái và Unicode (Thai script and Unicode) — James Đỗ Bá Phước

*Giới thiệu về chữ Thái và Font chữ Thái Sơn La (Introducing Thai letters and font of Son La) — Lò Mai Cương

*Font và Chương trình bàn phím chữ Thai (Font and keyboard for the Thai script) — Tô Trọng Đức and Phan Anh Dũng

Dự án bảo tồn tài liệu cổ chữ Thái (Project on preserving ancient Thai documents) — Đại diện sở VHTT Tỉnh Sơn La

*Công nghệ thông tin và bảo tồn văn hoá dân tộc (Developing projects on computerizing Thai culture and scripts) — Ngô Trung Việt and James Đỗ Bá Phước

Xây dựng dự án trang web chữ Thái (Website with Thai script) — Le Vinh Chien

Discussion

Several significant matters came out of the papers, the ensuing discussion, and the meeting with Cam Trong in Hanoi.

- The Tai people still value the traditional form of the script, and desire their children to learn it.
- Students who learn the Improved Script are unable to read the traditional script. Consequently, interest in the Improved Script appears to have cooled.
- The delegation from Son La proposed using the traditional Son La character set as a standard. They claimed that students who learn it can quickly learn other traditional dialects of the script. Considerable support was voiced for this proposal.
- No one at the workshop expressed any interest in the Tai Daeng script of Laos. In Hanoi, Cam Trong had expressed the belief that Tai Daeng is a distinct script from the Tai scripts of Vietnam. Given the fact that about 50% of the Tai Daeng characters are unique, and that the vowels carry a length contrast that is absent in the other dialects, I believe that Prof. Trong's position has considerable merit.
- I raised the issue as to the best spelling of the name of the script. One delegate said that it should be neither "Tai" nor "Thai", but rather "Tay", and that the use of a modifier before it, such as "Viet" or "Unified" is optional. The appeal of "Tay" is that it more closely matches their own pronunciation of their name for their language: In most

Romanizations, the “h” indicates aspiration, “ai” indicates a long /a:/, and “ay” indicates a short /a/. At supper that evening and the next day in Son La two or three others spoke to me in support of the “Tay” spelling, and no one spoke against it.

Decisions

Based on these discussions, Ngo Trung Viet, James Do, and I met in Hanoi on November 5 and made the following decisions:

- We will request the use of the spelling “Tay” if this does not conflict with other names.
- The character repertoire will be based on the Son La character set plus the following characters:
  - the aspirated consonants required for the Lai Chau dialects (3 hi/low pairs for /k^h/, /c^h/, and /p^h/)
  - consonants for writing “g” and “r” in Vietnamese loan words (2 hi/low pairs)
  - a vowel for writing Vietnamese â in loan words
  - Two sets of tone marks:
    - Spacing tone marks (similar to “e” and “j” in appearance)
    - Combining tone marks borrowed from Lao (MAI EK and MAI THO)
    - Usage of the latter pair is well established in the Tai community in the U.S. The former pair seems to be preferred in Vietnam. Inclusion of both pairs will create problems with searching and sorting which will have to be addressed. But they cannot be unified, as their combining class and storage order are different.

  The above repertoire constitutes the absolute minimum character set required for writing all the Tai dialects of Vietnam. It may be necessary to add additional characters later, as experience reveals the need. But it is felt that attempts to add other characters at this time, often involving characters of uncertain usage, would create questions which might take years to resolve and lead to prolonged delays.
- Storage will use visual order.

  Our initial intent was to use phonemic order. But further discussion since the workshop has revealed a significant advantage to visual order. Established keyboarding practices use visual order, as does the handwritten order. While input methods can be developed to support reordering of the input stream, those currently available are not advanced enough to provide a transparent editing environment after the input stream has been reordered. Experience gained by SIL with this script in the 1990s on a Macintosh based system revealed that the user experiences considerable confusion when both the input and output streams are reordered. Thus, visual order will result in a much improved user experience.

  Visual order is also used by the Lao script, to which the Tay is closely related.
- Sort order will be based on documents from Son La, if any are available; otherwise on Lao sort order.

- The encoding table will be patterned after the Lao Unicode block.
- Unicode character names will be patterned after the Lao Unicode names.

Other script dialects

The Son La character set on which we will base our proposal represents the central Tai Dam dialect of the script. Other Tai Dam regions show only minor variations from the Son La dialect. The following section is a preliminary examination of how the adoption of the Son La character set will impact other dialects.

Tai Don of Lai Chau Province and southern China

The Tai Don of Lai Chau have at least three distinct script dialects which depart from the Son La dialect. The script is also used in southern China. One sample that I have from China corresponds very closely to one of the Lai Chau dialects.

There are three types of differences between the Son La dialect and the Lai Chau dialects.

1. Characters which represent the aspirated consonants of Tai Don are not present in the Son La dialect of the script.
2. The various script dialects use different graphemes to represent the same sound.
3. Some characters of the Son La dialect occur in the other dialects with different phonetic values.

The Type 1 differences will be dealt with by the addition of the required characters to the Son La character set.

There is ongoing interest in Vietnam in developing a standard orthography for all of the Tai languages. It may be that the Son La character set will become that standard, and that the Type 2 and 3 differences will become moot. However, in the event that it is necessary to support the traditional Lai Chau dialects, there are two options for dealing with the Type 2 and 3 differences.

One option is to use regional fonts to express the differences as regional variations. That would work well for dealing with Type 2 differences in isolation. However, it can be shown that when the Type 2 differences are combined with Type 3, a situation is created in which the text is interpreted by the font. That is undesirable, though it may be tolerable if one is dealing only with historical documents.

The other option is to encode at least some of the Lai Chau variations as additional characters. I think this will be needed, at least to support the dialects of the script from southern China, where the people are less likely to adopt standards established in Vietnam.

Improved Alphabet

The Improved Alphabet was introduced in the 1950s and 1960s by Tai scholars in an attempt to standardize the script and correct its major deficiencies. Its future and the need to support it in Unicode are uncertain.

The Improved Alphabet differs from the Son La character set in two ways. One is in the variation of several consonant forms from the traditional forms. These can be treated as variant glyphs of the Son La characters.

The other variation is in the new vowel forms of the Improved Alphabet. The Improved Alphabet does not use any combining marks for vowels. New vowel forms are introduced which are spacing characters. These cannot be treated as variant glyphs of the Son La vowels, because the combining class is different. In the future, if a decision is made to support the Improved Alphabet, it will be necessary to add characters for the new spacing vowels.

Tai Daeng of Laos

Tai Daeng is spoken in both Laos and Vietnam, but the difference in their script from the Tai Dam seems to be more pronounced in the Laotian dialects. In the future, it may be necessary to encode Tai Daeng as a separate living script, as a separate historic script, or to add its character repertoire to that which is proposed above.

As noted above, some researchers have concluded that Tai Daeng writing, especially as used in Laos, is a distinct script from the Tai Dam and Tai Don scripts of Vietnam. The question is open to debate. On the one hand, about 50% of the Tai Daeng character set is identical to the traditional Tai Dam. Unifying the two would save some space in the Unicode code space.

On the other hand, there are factors that argue against unification. Tai Daeng writing shows length contrast on the vowels, whereas the Son La and Lai Chau dialects do not. Features which distinguish pairs of characters in Tai Daeng are sometimes blurred in the styles used in Son La and Lai Chau.

Unfortunately, our knowledge of Tai Daeng is very limited. I have had one reading primer from the 1970s and one undated hand-written text to work with. It is impossible to be certain as to the best course to take for Tai Daeng based on this limited information, but at this point I believe it should be encoded as a separate script.

Thai Song of central Thailand

As with the Tai scripts of Vietnam, there appear to be several dialects of Tai Song writing. Up to this point, I have found no indication that any of them are living scripts. The dialect that I have the best information on is very close to the traditional Tai Dam, or Son La, dialect. The writing style is very different, as sometimes expressed by giving radically different proportions to various features of a character. But, with two or three exceptions, the basic features which form each character are the same.

At this point, Thai Song should be considered a stylistic variation of Tai Dam. It may be necessary to revisit this question in the future, if contradictory information comes to light.