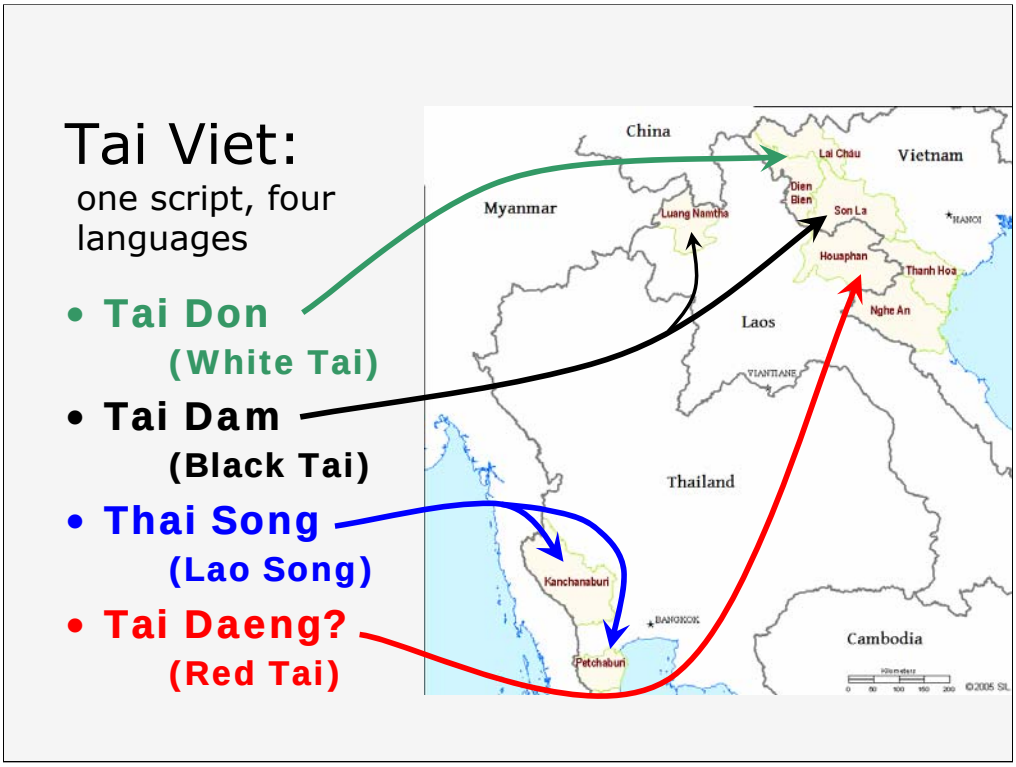
Unicode on the Front Lines

lessons learned
from encoding the
Tai Viet script

Over the last few years I have been involved in research on the Tai Viet script and in putting together two Unicode proposals for encoding the script, one of which was eventually accepted. I would like to share with you some of the lessons that I learned from that experience—especially for the benefit of anyone who may be in the process of writing their first Unicode proposal .



Tai Viet is the name of a script used by three or possibly four languages in SE Asia.

1. The two most prominent languages are Tai Dam and Tai Don, spoken in the northwestern provinces of Vietnam, and in neighboring regions of Laos and China.
2. The Thai Song people relocated from the Tai Dam homeland to central Thailand about 200 years ago. They have a language and script similar to the Tai Dam, but it is not known if the traditional script is still in use among them.
3. The Tai Daeng straddle the border of Laos and Vietnam. While their writing is similar to the Tai Dam and Tai Don, there are enough differences that it may be argued that it should be disunified. I will address that briefly at the end of my presentation.

- **Note** that *Tai Viet* is the name of the script, not of a language. The languages are the four mentioned here. However, sometimes for the sake of convenience, I may use *Tai Viet* to refer to the four languages as a group.

Tai Viet:
one script, four languages

- Also found in Australia, Canada, France, and the United States.
- Total population = 1 to 1.5 million.
- The script is being used in Vietnam and the United States.

•People speaking these languages can also be found in Australia, Canada, France, and the United States.

•The total population is 1 to 1.5 million.

•The script is being used at least in Vietnam and the United States. We have no information on its use in other areas.

Basic features

- Mostly monosyllabic
- Oral syllable pattern:
  - C_iV + tone
  - C_iVC_f + tone
- Written form complicated by
  - use of digraphs
  - non-linear vowel order

The Tai Viet languages are about 99% monosyllabic with simple syllable patterns (Ci + V + tone or Ci + V + Cf + tone).

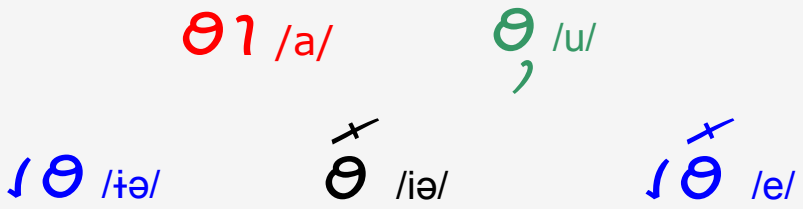
(The Ci is usually a single consonant. However, velar consonants may be labialized (i.e. they may form a cluster with /w/).

(Thai Song may also form initial clusters with /l/ and /r/)

In written form, the syllable structure is a little more complex, because some soundes are written with digraphs, and because the script has a non-linear vowel order.

Basic features

- Vowel marks are positioned before, after, above or below the syllable's initial consonant.



- As in other Tai scripts, vowel marks are positioned before, after, above, or below the syllable's initial consonant, depending on the vowel. Some vowels are written with digraphs that use a combination of two positions.

Basic Features

- Low and high series consonants

low series consonants are used for tones 1-3	 /pi¹/ ‘year’
high series consonants are used for tones 4-6	 /pi⁵/ ‘older sibling’

- Like most Tai scripts, Tai Viet has a low and high consonant series. Selection of a consonant from the high or low series is part of the system for marking tones.

Script Dialects

- Until 25 to 30 years ago…
  - Tai Viet was a hand-written script
  - There were no dictionaries or other documents to establish an orthographic standard

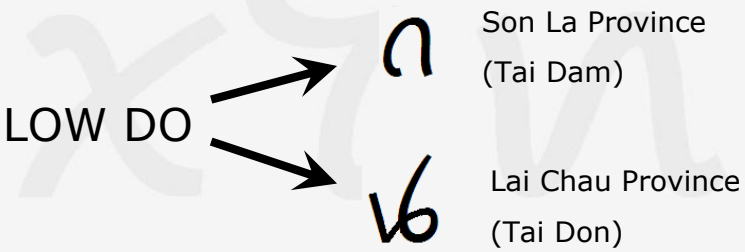
  ➡ **Many regional variations (script dialects)…**

Until 25 to 30 years ago, Tai Viet was only a hand written script. No dictionaries had been published. The result was a lack of any standard on how the script should be written.

•This led to many regional variations, or script dialects…

Script Dialects

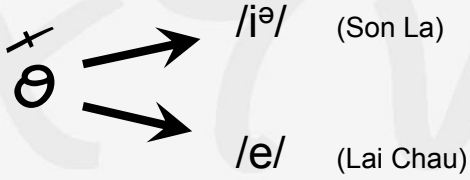
- Different regions may represent the same sound with different symbols:



Variations between dialects might show up as one sound being written by different symbols in different regions.

Script Dialects

- Different regions may use the same symbol to represent different sounds:



Or as one symbol being used to represent two different sounds.

Script Dialects

- The value of two symbols may be crossed between dialects:

	Son La	Mương Tăc
HIGH NGO	ꪆ	ꪜ
HIGH VO	ꪜ	ꪆ

In some cases, the sounds represented by two symbols may be crossed between dialects, as occurs here with the HIGH NGO and HIGH VO.

Script Dialects

- The value of several symbols may be skewed between dialects:

	Son La	Lai Chau
HIGH MO		
HIGH PO		
HIGH FO		

One of the more confusing situations is when the mapping between several symbols and several sounds are skewed. In this example, the Son La symbol for HIGH PO is used for HIGH MO in Lai Chau, the Son La symbol for HIGH FO is used for HIGH PO in Lai Chau, while a completely new symbol is introduced for the HIGH FO in Lai Chau.

Recent developments

- Tai script reform—1950s & 1960s
  - Attempted to reach agreement on a single uniform alphabet for use by all of the Tai people of Vietnam.
- Revived in the 1990s.
  - Desire to use the script in formal education

•Several developments in the latter half of the 20th century influenced the encoding of Tai Viet. Some of them grew out of a movement in the 1950s and 1960s by a group of Tai scholars in Vietnam to revise their own script. Their goal was to improve communication between the various Tai dialects of Vietnam by providing them with a single uniform alphabet that would be used by all of the dialects in Vietnam.

•The reform process moved slowly and languished in the 70s and 80s, but was revived again in the 1990s, along with a desire to use the script in formal education.

Tai Script Reform

- The reformers attempted to select a single symbol for each consonant.
- Most of the time they selected a symbol used by one dialect.

	Son La	Lai Chau
LOW SO		

The reformers attempted to compare the various dialects and select a single symbol for each of the consonants. Sometimes this involved selecting the traditional form of a consonant used by one of the dialects, and excluding the symbols used by the other dialects. For example, they selected the LOW SO used in Son La, and rejected the LOW SO used in Lai Chau.

Tai Script Reform

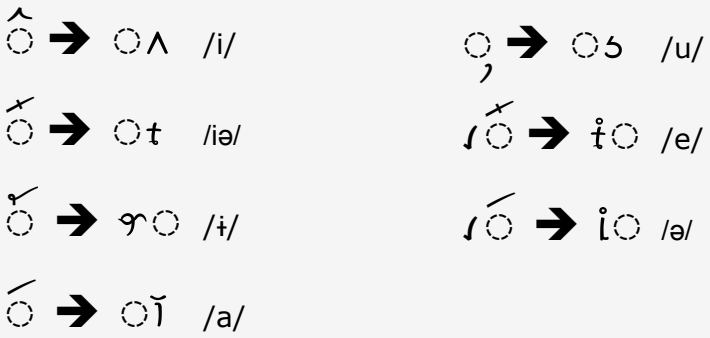
- In a few cases they introduced completely new symbols.

	Son La	Mương Tâc	
HIGH KO			

In a few cases they introduced completely new forms. In this example, they rejected all of the traditional forms of the HIGH KO, and introduced a new symbol for that character.

Tai Script Reform

- The reform replaced the combining vowels with spacing letters.



The most radical change made by the reformers was to the vowel system. About half of the vowels are combining marks written above or below the initial consonant. The reform movement decided to replace all of these combining vowels with new spacing vowels that sit on the base line.

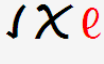
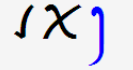
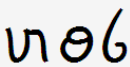
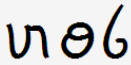
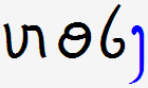
Tai Script Reform

- Not everyone accepted the reformed alphabet
- Result: another dialect was added to the script

Unfortunately, this reform process complicated an already confusing situation. Not everyone accepted the proposed reforms, so the reformed alphabet became in essence another dialect of the script that we had to deal with.

Other developments

- Tone marks
  - The traditional script lacked tone marks.
  - In Vietnam, MAI NUENG and MAI SONG were added for marking tone.

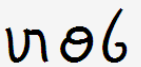
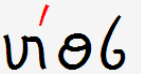
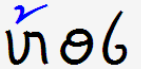
	unmarked	MAI NUENG	MAI SONG
low series consonant	 /sɨa¹/ ‘tiger’	 /sɨa²/ ‘mattress’	 /sɨa³/ ‘shirt’
high series consonant	 /hɔŋ⁴/ ‘under’	 /hɔŋ⁵/ ‘groove’	 /hɔŋ⁶/ ‘howl’

A less controversial change—the addition of tone marks—occurred spontaneously in different Tai communities at about the same time.

- One of the shortcomings of the traditional script was that tone was not written, beyond the selection of the high or low series for the initial consonant. This led to ambiguities in the text, which the reader had to resolve from context. To correct this, the Tai community in Vietnam introduced two tone marks, which they named *Mai Nueng* and *Mai Song.* These are spacing marks which are placed at the end of a word. When combined with the low and high series consonants, they are sufficient for marking all six tones that are used by these languages.

Other developments

- Tone marks
  - In the U.S., the Lao tone marks MAI EK and MAI THO were adopted.

	unmarked	MAI EK	MAI THO
low series consonant	 /sɨa¹/ ‘tiger’	 /sɨa²/ ‘mattress’	 /sɨa³/ ‘shirt’
high series consonant	 /hɔŋ⁴/ ‘under’	 /hɔŋ⁵/ ‘groove’	 /hɔŋ⁶/ ‘howl’

About the same time, the Tai community in the US adopted the Lao tone marks *Mai Ek* and *Mai Tho*. These are combining marks which are placed over the initial consonant. So there are now two competing sets of tone marks, and it has not been possible to unify them in the Unicode Standard, because they have different combining classes and occur in different positions in the data stream.

Questions

- Should we encode…
  - the reformed alphabet, or
  - the traditional script?
    - Which dialect of the traditional?
  - Or should we try to include all of them?

A year ago, we were still wrestling with these issues. Some wanted to use the reformed alphabet, others wanted the traditional form. But if we used the traditional form, which dialect should it be based on? Or do we try to accommodate all of the dialects?

Tai Viet Workshop

Fortunately, we got the answers we needed at a workshop on Preserving and Digitizing the Tai Viet Script, which was held in in Vietnam last November.

30 or 40 Tai delegates represented the various Tai speaking regions of Vietnam. They were joined by a handful of Vietnamese software developers, one Japanese, one Vietnamese-American, and myself.

Tai Viet Workshop

- There was strong opposition to the reformed alphabet.
  - Children who learned the reformed alphabet could not read the traditional.
- A proposal to use the Son La dialect was favorably received.

- The Tai delegates at the workshop were outspoken on one issue—they did not like the reformed alphabet. Literacy trials using the reformed alphabet had not produced acceptable results. Children who learned it could not read the traditional form used by their parents.
- A proposal was made that the traditional script of the Son La dialect should be used. Though no formal vote was taken, the discussion on the floor was heavily in favor of that proposal.

Tai Viet

- Current Status
  - Proposal for Tai Viet is in the Unicode pipeline
  - It is based on the Son La dialect
  - Three pairs of consonants added for Tai Don
- But…
  - Son La is a Tai Dam dialect.
  - Where does that leave the other three languages?

So where are we today?

•We have a proposal for Tai Viet script in the Unicode pipeline.

•It is based on the Son La dialect, but three pairs of consonants were added specifically for Tai Don.

But

•Son La is a Tai Dam dialect.

•Where does that leave the other three languages?

The future—Tai Don

- Additions for Tai Don may be needed…
  - if the Tai Don people reject the Son La dialect as a standard
    - Tai Don in China have not been involved in these discussions.
  - to record historical documents which use the old Tai Don forms.

1. The repertoire that is in the pipeline is adequate for writing Tai Don, providing the Tai Don people accept the Son La dialect as a standard.
2. If they do not, it may be necessary to add additional characters for that language.
   Note that the Tai Don in China have had no input on the current proposal.
   There is also the question of whether the old Tai Don forms may be needed for recording historical documents.

The future—Thai Song

- Vowel length is not marked in written data
- Similar to Tai Dam with strong stylistic variation
- Appears to be of only historical interest
- A Thai Song style font with the currently proposed encoding should be sufficient

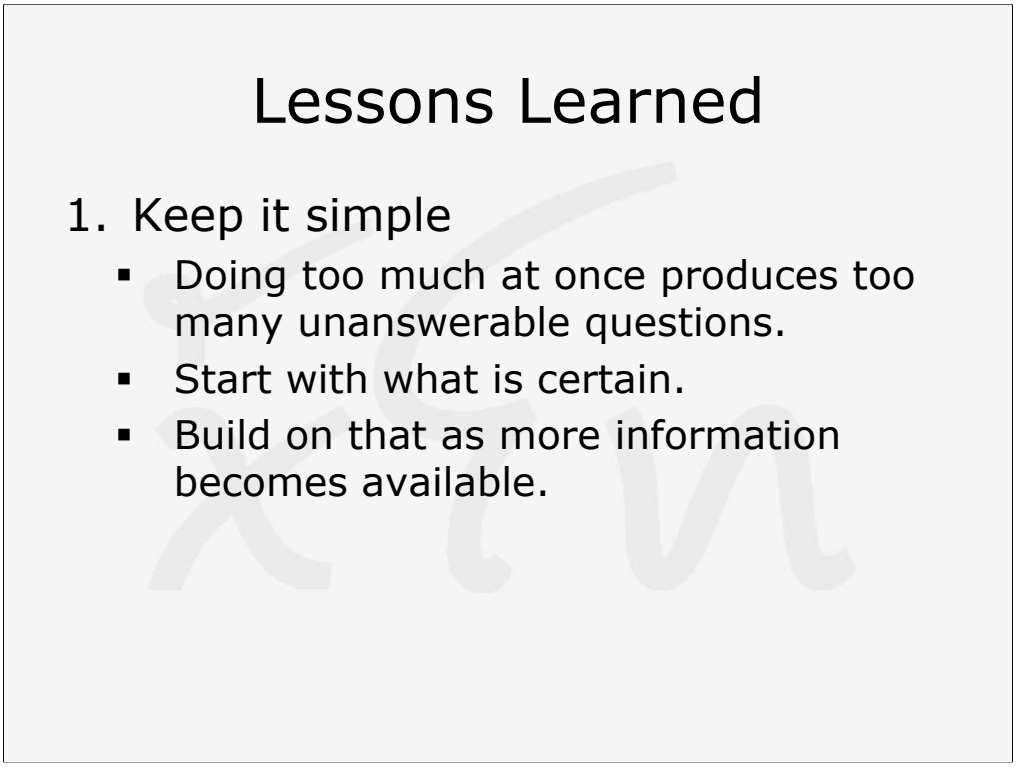
1. With regards to the Thai Song, the most significant linguistic difference between Thai Song and Tai Dam is that the former has length contrast on all the vowels, whereas the latter has length contrast only on /a/. One might expect that to show up in the written form of the Tai Song, but the limited data that I have do not show any length contrast.
2. What I do find in the data is that the basic letter forms tend to correspond to the Son La dialect of the script, but they have a very strong stylistic variation.
3. In addition, I have not seen any evidence that the script is currently being used among the Thai Song. So it is primarily of interest for historical purposes.

Under these circumstances, it should be sufficient to use the current Tai Viet encoding with a font based on the Thai Song style—unless further information comes to light.

The future—Tai Daeng

- 50% of consonant forms are unique
- 70% of vowel forms are unique
  - Tai Daeng has length contrast on all vowels
  - Tai Dam and Tai Don have length contrast only on /a/
- Tai Daeng should be encoded as a separate script.

1. 50% of Tai Daeng consonant forms are unique
2. 70% of vowel forms are unique—largely because of the need to mark vowel length on all vowels in Tai Daeng.
3. Even though it has a superficial resemblance to Tai Viet, I believe Tai Daeng should be encoded as a separate script.



I have shown you some of the issues that we had to deal with to encode the Tai Viet script. In the process, I made a few mistakes, and learned a few lessons. I'd like to share some of the lessons with you.

1. Keep it simple

The many dialects were a stumbling block for me when I started. My assumption was that a single massive comprehensive proposal that included the reformed alphabet and all of the dialects would produce more uniform results than going about it piece meal. That proved to be a delusion.

Trying to do it all at once produced too many unanswerable questions, which only delayed the process.

When working with a complex proposal, start with what is certain. Build on that as more information becomes available.

Lessons Learned

2. Encoding standard ≠ orthography standard
   - Orthography = how words are spelled (i.e. what symbols are used)
   - Encoding = what are the abstract characters and how are they encoded
   - Know which questions are orthographic, and which are encoding

When dealing with a writing system that is in flux, one must be aware of the difference between an encoding standard and an orthographic standard.

- Orthography defines how symbols are arranged to spell words.
- Encoding identifies the abstract characters of the orthography and defines how they are encoded.

Having to deal with both orthographic questions and encoding questions at the same time adds a lot of complexity to the project. But at least it helps if you are aware of which questions are orthographic issues, and which are encoding issues.

Relationships are key

3. Contact with the user community is essential
   - Social & political factors influence orthography, which influences encoding
4. Communicate with the UTC
   - Solicit their feedback and guidance

•*Relationships are key - contact with the user community is essential*

In theory, Unicode is non-political. In reality, at the very least, orthography is influenced by social and political factors. That in turn influences the encoding that becomes part of Unicode.

I had been in contact with the Tai community in the US for many years. This was very valuable, but it did not allow me to keep abreast of developments in the homeland. About two years ago, Debbie Anderson put me in contact with Ngo Trung Viet in Hanoi, and it was the relationships which developed from that contact which enabled the successful completion of the proposal.

•*Relationships are key – communicate with the UTC*

If you are writing your first Unicode proposal, especially if it is for a complete script, one of the things you will discover is that there is more detail involved than you ever imagined. And as far as I know, there is no one document that tells you what all the pieces are. Members of the UTC can guide you through the process.

One of my weaknesses is that I want everything to be perfect before I show it to anyone. Consequently, by the time I do show it to someone, it's too late to get the amount of feedback I need. Avoid my mistake, and get all the feedback you can as early as possible.

•**So, with these points in mind…**

Go write a Unicode proposal!