Computing in Vietnamese: Progress & Challenges

{James} ĐỖ Bá Phước
杜 伯 福

IMUG 2005-05-19

Overview

- Vietnamese writing
  - Latin: Quốc ngữ
  - Ideographic: Chữ Nôm
- Considerations
  - Repertoire
  - Character encoding
  - Input methods
  - Fonts

Quốc ngữ
{National script}

Orthographic units

- Vowels

  a ă â e ê i o ô ơ u ư y

- Consonants

  b c d đ g h k l m n p q r s t v x

- Tone marks

  ◌̀ ◌̉ ◌̃ ◌́ ◌̣

- A vowel can combine with one tone mark

192 characters, 6 “too many”

- A total of 192 upper- and lower-case “pre-composed” characters
- 6 characters beyond 8-bit character set
  - 6 characters missing from original ISO 10646 repertoire (in 1988)
  - Restored through Unicode-10646 merger

43 8-bit character sets

- Pre-composed
  - Dual fonts
  - TCVN 5712:1995, aka “ABC”
    - (TCVN = Tiêu chuẩn Việt Nam {Vietnam Standard})
- Glyph overlap
  - VNI
- Combining
  - Windows Vietnamese (cp-1258)

Unicode

- Encodes both:
  - Combining characters (from Unicode)
  - Pre-composed characters (from ISO)
    - Getting more widely supported
- Wide acceptance after and for the Web
- TCVN 6909:2001
  - Pre-composed characters only

Vietnamese writing

- Handwriting
  - v i e ^ t ´ ⇨ viết {write}
  - v i e ^ ´ t ⇨ viết {write}
- Telex
- Typewriter
  - Dead-key
    - Carriage stops at combination of diacritics, then
    - Base letter
- Computer

Telex convention

- a ă â e ê i o ô ơ u ư y

  aw aa ee oo ow uw
- b c d đ g h k l m n p q r s t v x

  dd
- ò ỏ õ ó ọ

  f r x s j
- Example: vieets ⇨ **viết**

Computer input methods

- Telex
- VNI
- VIQR (VIetnamese Quoted-Readable)
  - Mnemonic
  - Internet RFC 1456
- TCVN 6064:1995
  - Orthographic units

VIQR, aka VietNet

- a ă â e ê i o ô ơ u ư y
  a(a^ e^ o^ o+ u+
- b c d đ g h k l m n p q r s t v x
  dd
- 
  ` ? ~ ’ .
- Example: vie^’t ⇨ **viết**

Are diacritics necessary?

- Ma ⇨ ghost
- Mà ⇨ but
- Mả ⇨ tomb
- Mã ⇨ code
- Má ⇨ mother
- Mạ ⇨ rice seedling
- Mua ⇨ to buy
- Mưa ⇨ rain

TCVN 6064:1995

- Corresponds to orthographic units
  - Closer to handwriting
- Native to:
  - Windows XP
    - Emits combining character sequences
  - Mac OS X
    - Emits pre-composed characters in OS X 10.4 Tiger
    - Was emitting combining character sequences

TCVN 6064:1995

`	ă	â	ê	ô	◌̀	◌̉	◌̃	◌́	◌̣	đ	-	◌̲đ	BkSp	
Tab	q	w	e	r	t	y	u	i	o	p	ư	ơ		
CapsLock	a	s	d	f	g	h	j	k	l	;	'	\	Enter	
Shift	z	x	c	v	b	n	m	,	.	/	Shift			
Control	Alt										AltGr	Control		

`	1	2	3	4	5	6	7	8	9	0	-	=	BkSp	
Tab	q	w	e	r	t	y	u	i	o	p	[	]		
CapsLock	a	s	d	f	g	h	j	k	l	;	'	\	Enter	
Shift	z	x	c	v	b	n	m	,	.	/	Shift			
Control	Alt										AltGr	Control		

Input software

- http://unikey.sourceforge.net
  - Different input conventions
  - Multiple character encodings
  - Clipboard converter
    - Extremely convenient for handling documents in legacy encodings
  - Free, lightweight, powerful
    - Use from your USB memory device!

Localization

- Locale
- Translation of computer terminology
- Windows XP and Office 2003 SE
  - Using LIP (Language Interface Pack)

Progress

- Vietnamese web sites are universally in Unicode
  - Exception: http://www.vietmercury.com
    - This site never shows up in Vietnamese web searches!
- Search
  - Web: Google, Yahoo!, MSN
  - Desktop: Google, MSN
- Blogs, wikis, ...
- Desktop & server applications

Progress & Challenges

- Unicode-savvy?
  - Yahoo!Mail, AOL, AIM Mail: charset “iso8859-1”
  - Eudora
- Unicode-savvy!
  - Outlook Express, Outlook, Thunderbird
  - Gmail, Netscape Mail
  - Yahoo!Messenger, MSN Messenger, Skype
- Not enough Unicode fonts with Vietnamese
  - Example: Trebuchet MS (which reverts to Arial)

Challenges

- User education
  - GIGO
    - Any non-Unicode string
- Legacy
  - Encodings
    - Remove all non-Unicode fonts
- No physical standard keyboard

Chữ Nôm
{Demotic script}

Chữ Nôm 字喃

- Started to appear in the Xth century, after a thousand years of Chinese rule
- Based on Chinese characters
- In use for the next thousand years
- Now replaced by Quốc ngữ

Nôm example

<table>
<tr><th colspan="3">Vietnamese</th><th>English</th></tr>
<tr><td>Quốc ngữ
Latin script</td><td colspan="2">Nôm
ideographic script</td><td>Latin script</td></tr>
<tr><td rowspan="2">cột</td><td colspan="2">楫</td><td rowspan="2">pillar</td></tr>
<tr><td>Meaning
mộc {tree}</td><td>Sound
cốt {sound}</td></tr>
</table>

Nôm dictionaries

- 1971 ~ *Tự điển chữ Nôm {Nôm Dictionary}*, Nguyễn Quang Xỹ & Vũ Văn Kính
- 1988 ~ *Chu-Nomu Jiten* 字喃字典, Takeuchi Yonosuke
- 1999 ~ *Đại từ điển chữ Nôm {Nôm Super-Dictionary}*, Vũ Văn Kính
- 2004 ~ *Giúp đọc Nôm và Hán-Việt {Nôm & Hán-Việt Reading Guide}*, Father Anthony Trần Văn Kiệm
- Soon ~ *Từ điển chữ Nôm tiếng Việt {Vietnamese Nôm Dictionary}*, Nguyễn Quang Hồng

Nôm standard encoding

- First proposed in 1992
- Unicode 3.1
  - 9,299 characters, of which
    - 5,067 characters in BMP (CJKV Extension A)
    - 4,232 “Nôm proper” in Plane 2 (CJKV Extension B)
- IRG: CJKV Extension C
  - About 2,200 additional characters
  - (IRG = Ideographic Rapporteur Group)
  - (CJKV = Chinese, Japanese, Korean, Vietnamese)

Nôm input methods

- http://www.viethoc.com/hannom/bango_intro.php
  - Source
    - 6 dictionaries
    - Other databases
  - Currently available for input
    - 16,638 Chinese characters (from 22,975 possible)
    - 11,600 Nôm characters (from 20,732 possible)
- HanoKey
- HanoSoft

Nôm fonts

- 9,299 characters
  - Mojikyo Institute, Tokyo
  - Dynalab, Taipei: DFSong Light Vietnam
- 30,000+ characters
  - Đạo Uyển, Viên Chiếu Monastery: HanNom A & B
- 17,000+ characters
  - Nôm Na Group, Hà Nội: Nôm Na Tong Light
  - 4,415 basic Hán-Nôm components

Nôm online

- http://www.viethoc.com/hannom/tdnom_beta.php
  - Nôm Annotated Dictionary
  - Uses HanNom fonts, Java applet
- http://nomfoundation.org/nomdb/lookup.php
  - Nôm Lookup Tool
  - Uses .GIF images (was SVG)
- http://www.huesoft.com.vn/hannom/
- http://sager-pc.cs.nyu.edu/~huesoft/
  - Việt-Hán-Nôm Dictionary

Challenges

- Repertoire
  - Newly discovered characters
  - No coordination between active groups
- Character encodings
  - Slow international standardization process
  - Private Use Area
  - No coordination between active groups

Challenges

- Input methods
  - Large character repertoire
- Fonts
  - Unicode surrogates

Other options

- Presentation
  - SVG (Scalable Vector Graphics)
  - CDL (Character Description Language)
    - http://www.wenlin.com/cdl/

Wrapup

Brief history

- Xth century ~ first Nôm writing
- 1651 ~ first Latin-based dictionary
- 1910 ~ Quốc ngữ adopted nationally
- 1991 ~ Quốc ngữ orthographic units in Unicode 1.0
- 1993 ~ RFC 1456 (VIQR)
- 1993 ~ Quốc ngữ pre-composed in Unicode 1.1, ISO/IEC 10646-1
- 1995 ~ TCVN 5712 (8-bit), 5773 (Chữ Nôm)
- 1995 ~ TCVN 6064 (keyboard)
- 2000 ~ Chữ Nôm in Unicode 3.1
- 2001 ~ TCVN 6909 (Unicode)
- 2004 ~ First International Nôm Conference
- 2005 ~ Vietnamese Windows XP and Office 2003 SE

Dichotomies (& synergy?)

- Latin, ideographic
- Different encodings
- Combining, pre-composed
- Telex, VNI, TCVN 6064
- Fonts
- Active working groups

Challenges

- Standardization
  - Repertoire
  - Character encoding
  - Input methods
  - Fonts

	Qn	Nôm
Repertoire		
Character encoding		
Input methods		
Fonts		

Challenges

- Usage
  - Legacy
  - Application support
  - User education

	Qn	Nôm
Legacy	☹	☺
Application support	☺	☹
User education	😐	😐

Ultimately

- To make Vietnamese like any other language (such as English) in computers
- Goal: an ordinary user of Vietnamese on computers should not have to know about UTF-8 or character encodings at all

Thanks!

jdo@pacificlinks.org
http://vietual.blogspot.com

Acknowledgements

- With thanks to:
  - Roger Sherman
  - David Murphy
  - James Turley
    - for enabling the presentation at IMUG and on the web
  - Hồ Văn Tiến
  - Ngô Thanh Nhàn
  - Ken Lunde
  - Tex Texin
    - for comments and corrections
  - The IMUG (International Macintosh Users Group) audience
    - for interesting questions and a very lively exchange

Q & A